2017, Viviana Daniels Coneo & Elkin Jose Villanueva Niebles

HOW THE USE OF TEST SPECIFICATIONS AFFECTS TEST DESIGN IN TERMS OF THE SELECTION OF LEVEL APPROPRIATE TEXTS AND THE CREATION OF LEVEL APPROPRIATE QUESTIONS IN AN EFL PROGRAM

BY

VIVIANA DANIELS CONEO
ELKIN JOSÉ VILLANUEVA NIEBLES

Graduate Specialist, Universidad del Norte, 2017

RESEARCH STUDY

Submitted in partial fulfillment of the requirements
for the degree of MASTER OF ARTS IN ENGLISH LANGUAGE TEACHING of Universidad del Norte, 2017

Barranquilla, Atlántico
Colombia

Master's Supervisor
Erica Ferrer Ariza, M.A.

AFFIDAVIT

I, Viviana Daniels Coneo, hereby declare that this master's thesis has not been previously presented as a degree requirement, either in the same style or with variations, in this or any other university.

(add digital signature here)

VIVIANA DANIELS CONEO

AFFIDAVIT

I, Elkin José Villanueva Niebles, hereby declare that this master's thesis has not been previously presented as a degree requirement, either in the same style or with variations, in this or any other university.

(add digital signature here)

ELKIN JOSE VILLANUEVA NIEBLES

Abstract

This paper consists of analyzing whether the use of Test Specifications affect test design in terms of the selection of level appropriate texts and the creation of level appropriate questions in an EFL program. This research aims at analyzing the language level in the texts and the quality of items prior and after the implementation of Test Specifications. The methodology used to conduct this research is Mixed Research due to the fact that both Qualitative and Quantitative approaches are integrated, and Qualitative aspects can be explained in a Quantitative way. The research is carried out revising eighteen samples of listening and reading tests of levels two and four, prior and after the implementation of Test Specifications, of an EFL program at a private university in Colombia. The main instrument used to make the revision is a checklist, as a manner of examining validity, text language appropriacy and test items quality. The results obtained show that Tests Specifications are well designed and relevant but when put into practice, they might not be well executed. Also, there are some aspects that need further explanation. For example, the approximate number of questions exams should have taking into account that the time frame is fifty minutes. A general conclusion is that documents such as Test Specifications need to be supported with complementary work since its sole creation does not guarantee significant change. The process of Test Specs creation, implementation, and improvement should be continuous to be able to observe changes in assessment tools design.

Acknowledgments

First off, I want to thank God who has given me the strength to keep going. Thanks to my loving husband, Luis, for all his support and patience, my beautiful two sons, Santiago and Samuel, I want you to be proud of your mom. I love you guys with all my heart and soul. To the rest of my family, especially to my mother who I know she is up there very proud of me, I love you and I miss you so much. To my thesis partner, Elkin, who has become a friend throughout this process. Last but not least, to our tutor, Erica, thanks for all the support you have given us.

Viviana Daniels Coneo

First of all, I want to thank God because that is who I look up to. He has graced my life with opportunities that I know are not of my hand or any other human hand. He has taught me that gratitude reciprocates. Thanks to my Family, who are the three people in my life that I want to make the most proud of me: To my mom, Duvit, who has shown me that Education and commitment are good paths to progress in life. To my dad and sister, Heberto and Eliana, from whom I learned to keep a balance and not to take life too seriously. To my two other mommies that are in heaven, Nancy and Genoveva, who taught me how to work honestly, how to listen and be empathetic with others. More than words for everything you all gave to me. To the rest of my family for loving and taking care of us when we were growing up. To my thesis partner, Viviana, a friend who constantly encouraged me to keep moving forward in this paper. To my tutor and role model, Erika, for all the patience and efficient support she had with each one of us. Thank you all.

Elkin Jose Villanueva Niebles

Table of Contents

List of Tables

Table

List of Figures

Figure

1. Introduction

Research is a crucial aspect of academic life. Human beings, by nature, seek to provide explanations to complex situations and events that occur in reality. Throughout time, there has been a continuous concern about illustrating the exact picture of material world, and researchers do their best to discover truth based on previously constructed knowledge, theories, and studies. In that line of thought, Burns (2008) argues that "research only matters if you do research that matters" (p.12). Thus, doing quality research is fundamental in people's lives due to the fact that asking worthy questions not only requires replying worthy answers but also benefiting and making significant contributions to the globalized world. Likewise, if researchers do authentic research that demands effort, explorations, and questions that are worth answering, they will likely obtain valuable and interesting results that may trigger innovation and educational progress.

However, when it comes to educational research, it seems that the focus of research tends to be on some repetitive aspects. White (2009) states that researchers unfortunately tend to concentrate more on reflecting about teaching methods and approaches, and they tend to forget about studying assessment as another way of measuring learning. In other words, White says that, in research, assessment is as essential as teaching practices because teachers can become better educators if they know how to effectively assess students. Therefore, reinforcing research on assessment can be also significant to contribute to the improvement of teaching and learning and not just implementing assessment as a measuring matter.

But the word assessment is still seen by many as a daunting threat that provokes anxiety instead of positive feelings. Students often associate assessment with grades and failure rather than to motivating learning experience that provides valuable feedback. Thereby, this concept of

assessment has to be dealt with carefully in education due to the fact that Assessment is not the same as testing. According to Brown (2004), testing is a method of measuring a person's skills, knowledge, or performance in a given field while assessment is the continuous process that includes a wider domain. This means, in other words, that testing is just an instrument of assessment that contributes to measure certain teaching and learning aspects.

Similarly, Bulter & McMunn (2006) see assessment "not as a thing that is done to students, but as a process that can lead to improve learning" (p.2). In this way, assessment not only allows teachers and/or programs to collect information but also identify students' difficulties, monitor progress, and provide appropriate feedback.

In relation to tests as part of assessment, Brown (2004) presents five principles of assessment to effectively design tests. These principles are practicality, reliability, validity, authenticity, and wash back. They are specially applied in classroom assessment designs to see how effective, accurate, useful, or down to earth a test can be. Educational staff should be informed about these principles and the benefits they have.

Moreover, some institutions are using more and more these principles when designing a test by following certain elaborated guidelines. These guidelines are usually known as *Test Specifications* or *Specs*. Fulcher (2007) says that "test specifications are generative explanatory documents for the creation of test task" (p. 52). Hence, Specs describe step by step how to select, arrange, and separate test items, how to structure test layout and locate passages, the way questions are formulated, the manner of scoring, and so forth. In other words, Specs are the outline of the test, how the test looks like in terms of content and appearance. Likewise, Test Specifications represent an important role in education because they can contribute to an effective test design and, therefore, improve the quality of inferences made about students'

language proficiency from evaluation results. Test Specifications also known as Specs are crucial if Language programs decide to standardize their own assessment process. Teachers can be great contributors if they know how to properly design assessment tests by following certain Specs.

For Davidson and Lynch (2002), the initial purpose of Test Specifications is to have the same assurance of reliability and validity in a test. In this way, Specs can have a broader and more compact impact not only on test development but also on teaching and learning practices. Thus, considering the importance of test specifications in producing valid assessments, the purpose of this research is to explore the role that the use of test specifications play on classroom test design in terms of the selection of level appropriate texts and the creation of level appropriate questions in an EFL program. This general objective motivates other specific objectives such as the analysis of the language level in the texts prior and after the implementation of test specs and the appropriacy of questions used prior and after the implementation of test specs.

Classroom assessment is a very important area to explore and develop further in Colombia as argued by Lopez and Bernal (2009), who indicate that assessment is not even included in the plan of study of English teaching undergraduate programs. Assessment is an area that ideally informs teaching and learning. It provides information that is used for the purpose of making decisions; therefore, it is an area that should not be taken for granted but rather explored and strengthened. Working on areas that allow for the design of more valid assessment instruments could help teachers and institutions develop an awareness of assessment literacy.

This paper is divided into eight chapters. Introduction is the first chapter that explains the concept of assessment and the importance of specs in research. The second chapter refers to the rational of the paper, it includes the context in which the research is carried out and the question

to be answered. Chapter Three presents the general and specific objectives of the paper. Chapter four deals with theoretical framework that includes a detailed definition of the most relevant concepts used in the paper such as *assessment*, *classroom assessment*, *assessment literacy*, *testing*, and *test specifications*. The fifth chapter shows Methodology which includes a general characterization of the approach to research, the study followed and its justification; it also elaborates on the decisions made in terms of research design, actions to collect data; instruments or tools to collect, and procedures or steps followed. This section does not only remind us the research question but also describes formal reading and listening assessment tools used as the material -focus of analysis for this research project, and the instruments to be implemented in the data collection process. Chapter six (results) describes a detailed and organized analysis of the data collected. The findings and the analysis of the results are presented in the seventh chapter called Discussion and conclusions. After the application of instruments, the main results of the study are described in this chapter and supporting evidence is illustrated. Finally, references are listed, followed by the appendixes, which contain all the instruments used in the data collection process to illustrate evidence of the work done.

2. Rational

In the globalized world that we currently live in, English is an essential part of our social and professional lives. Having this in mind, governments around the world are investing important financial resources and effort to try to equip their people with this relevant communication tool. Colombia is no exception in this matter. In the "Ley General de Educación" (Ley 115 de 1994), English is stated by the government as one of the mandatory and essential areas in elementary school and higher education.

Education, as we know it, entails different processes and different actors. One of the most important processes in education is assessment since it provides valuable feedback when taken seriously and all the actors involved in the process are fully aware of its relevance. Influenced by this, we decided to carry out this research to see what the influence of a document, called "Test Specifications" can have (if any) on the design and the decisions made by the teacher-coordinators in charge of designing the reading and listening exams in a private institution in Barranquilla.

We have decided to carry out this research at the Foreign Language Department in Universidad del Norte, a private institution in Barranquilla, Colombia. The Foreign Language Department offers different foreign languages options: English, French, German, Italian, Portuguese, Mandarin, Japanese, and Spanish for foreigners. It is divided into two major areas: Extension Program and Undergraduate Program. In the Extension Program, there are all the different programs the Language department offers to the community that does not study a major in the University, including preparation courses for international exams, English for Schools, English for Companies, among others. The Undergraduate Programs area caters for the foreign language requirement of the undergraduate student population through credit-bearing and non-

credit bearing programs. This research will focus on two levels of the ELP (English Language Program), which is one of the non-credit bearing programs offered by the university.

Most of the programs have eight levels and include a preparatory 80-hour course for students who begin the programs at A0, A1- level according to the Common European Framework for Reference (CEFR), as informed by the placement exam. Each one of the levels in the program has a level coordinator, who is a teacher that has taught the level and is familiarized with it. This teacher has a leading role and among his/her functions has to have regular meetings with the teachers teaching in his/her level, observe their classes, according to a pre-established plan, once a semester, and design the exams all the students in the level will take. The number of students taking the exams the level coordinator designs can reach up to six hundred and the number of teachers applying these exams ranges from eight to fifteen. Students' grades are normally followed up on and inferences about students' achievement, inter rater reliability, and program effectiveness are made on the basis of these results. This is the reason why designing exams of high quality is of paramount importance in the institution.

In each one of these levels, the development of basic skills (reading, listening, writing, and speaking) is a priority. Special emphasis is made also on Grammar and Vocabulary. Therefore, to make assessment consistent with the skills taught, there are five formal assessment instruments or tests in place, each focusing one of the skills, and a cumulative final exam. All the tests are designed by the level coordinators and analyzed, improved, and approved by an assessment coordinator.

All teachers know that designing assessments is not an easy task; reason why White (2009) makes an important emphasis on the term assessment literacy, which he defines as "an understanding of the principles of sound assessment" (p. 5). As it was mentioned in the previous

section, the principles of assessment suggested by Brown (2004) are practicality, reliability, validity, authenticity, and washback. All these principles will be further explained in detail.

At a certain point in our teaching practice, we all have been assessment designers, and as such, we have faced difficulties regarding the decision making process about the language level of the texts we choose to include in the assessment and the appropriacy of the questions we make in the assessment. This is the main reason why the English Department of Universidad del Norte has decided to create test specifications for all the levels in the Undergraduate Program.

The project of the creation of test specifications began with a piloting process in the levels two and four. Test specs are being constructed and improved in these levels, but before the project moves on to the other levels, this research proposes to know if having these documents may actually influence the assessment design. This situation motivates the following question:

How does the use of test specifications affect test design in terms of the selection of level appropriate texts and the creation of level appropriate questions in an EFL program?

According to White (2009) "assessment is a central element in the overall quality of teaching and learning in higher education." He also points out that "assessment is probably the most important thing we can do to help our students learn." (p. 5) Taking this into account, assessment could be considered as one of the most important elements in the teaching and learning process. Therefore, having a document that can help teachers make better decisions could be beneficial for their practice. Consequently, establishing whether having test specifications in place for a program make any difference in test design will help the Foreign

Language Department to determine if test specs are useful to assessments designers or they still need improvement in order to be a useful tool to design have valid, fair, and high quality assessments.

3. Objectives

As it was mentioned in the introduction, in this section, the objectives on which this research will be focused are described.

3.1 General Objectives

3.1.1 Explore the role that the use of test specifications play on test design in terms of the selection of level appropriate texts and the creation of level appropriate questions in an EFL program.

3.2 Specific Objectives

3.2.1 Analyze the language level in the texts used prior and after implementation of test specifications.

3.2.2. Analyze the quality of items prior and after the implementation of test specifications.

4. Theoretical Framework

The aim of this section is to provide a theoretical framework for this research. We will present some theories that support the umbrella concept for this research, which is assessment. We will see in detail what the differences between assessment and testing are, the principles good assessment should be based on, the different kinds of assessment, and test specifications' benefits and possible constraints. We will also revise other studies carried out on test specifications in Colombia and abroad.

Since the beginning of times people have been concerned about what surrounds them; this is the reason why they try to understand the environment. "The means by which they set out to achieve these ends may be classified into three broad categories: experience, reasoning, and research" (Cohen, et al, 2013, p. 3). This first notion of research suggests that this is not a new interest of people, since long time ago people have been inquiring about the things around them, and this can be done formally or informally.

Taking into account the purpose of this research, which is to explore the role that the use of Test Specifications play on test design in terms of the selection of level appropriate texts and the creation of level appropriate questions in an EFL program, it is worth establishing the meaning of the term assessment and its difference with the term testing. Assessment is "a process of reasoning and evidence gathering carried out in order for inferences to be made about individuals" and the "task of establishing the meaningfulness and the defensibility of those inferences as being the primary task of assessment development and research" (McNamara & Roever, 2006 p. 12)

Language assessment is an area of which some points are worth highlighting. Bachman (2004) states that language testing has "the potential for helping us collect useful information

that will benefit a wide variety of individuals;" in his opinion, "scores from language tests are used to make inferences about individuals' language ability and to inform decisions we make about those individuals" (p. 3). To have a clear notion of how to use and interpret the results gotten from assessments is of paramount importance because, otherwise, it would be a worthless process if no changes and improvements in the programs are made based on results.

Bachman (2004) establishes as one of the many uses assessment has the fact that it helps "to make inferences about abilities or attributes such as lexical knowledge, sociolinguistic awareness, language aptitude, or motivational orientation" (p. 9). Results of assessments can offer information on the characteristics of students as their strengths and weaknesses, their success in a language course, or how skillful they are in the language they are studying.

As mentioned in previous chapters, assessing should not be seen as a burden, but as a way to get essential feedback not only of the students' progress but also of the process carried out by the instructor. Assessment is sometimes a term people do not understand well and can be confused with the term test. According to Brown (2004), "in educational practice, assessment is an ongoing process that encompasses a wide range of methodological techniques" (p. 3). In his book, it is established that assessment and test are different since the former is a process and the latter is defined as a "method of measuring a person's ability, knowledge, or performance in a given domain." (p. 4) Tests are also seen by these authors as a genre of assessment techniques; they are part of administrative procedures and students must take them at specific times in a curriculum knowing that their answers will be measured and evaluated.

Flores (2016) states that classroom assessment deals with all "the formal and informal activities that take place in the classroom, developed, used, or selected by the teacher according to the teaching context, the instructional goals, and the teacher's knowledge of their students on

aspects such as the students' progress, learning styles, or strengths and weaknesses"(p. 7). The term classroom assessment is relevant at this point because it is the focus of this study, the classroom is the context in which all this process takes place. As Flores (2016) states, all the assessment activities done in the classroom (quizzes, classroom conversations, assignments) have to be thoughtful and have a clear purpose. They should not be deliberate decisions we make on a rush.

Based on these definitions of the term assessment, it is noticeable that, more than anything, it is seen as a process and not as a finished product. As a process, it entails "the act of collecting information about individuals or groups of individuals in order to better understand them. The twin purposes of assessment are to provide feedback to students and to serve as a diagnostic and monitoring tool for instruction" (Butler & McMunn, 2006, p. 2).

From this, it can be inferred that assessment is not something instructors or teachers do to their students but a process that makes possible the improvement in their learning. As it was mentioned before, there has to be a clear difference between assessing and testing students. Following the same line of ideas, tests, are a category of assessment, a genre of assessment techniques. They are part of administrative procedures, done on times set up since the beginning of a school year, a term or a course, being aware of the fact that the responses given by test takers are being measured and evaluated.

Another important term that needs to be tackled at this point is assessment literacy, an important term that teachers should have clear, which consists on having high levels of awareness of knowing what, why, when, and how to assess. Webb (as cited in White, 2009) affirms that "assessment literacy is the knowledge about how to assess what students know and can do, interpret the rules of these assessments, and apply these results to improve students'

learning and program effectively" (p.7). So, a teacher's level of assessment literacy can influence students' learning and course achievement. Assessment literacy is important because it provides learners affordances to self-monitor- rehearse, practice, and receive feedback; it helps and improves learning (Webb, 2002 in White, 2009). The way assessment is implemented has changed the nature of teaching and learning. Currently, some teachers might see assessment as a formative process rather than summative when it follows certain standards that might contribute significant feedback for students and teachers.

Bulter (2006) also classifies assessment into formative and summative. He states that summative assessment focuses on scoring events that were previously studied in a teacher's grade book; these events are usually evaluated at the end of the lessons and report students' achievement. On the contrary, formative assessment establishes certain objectives for students to accomplish in class and provides progressive feedback to them. For example, teachers use formative assessment on a daily basis in their classroom because it fosters more progress, and they use summative assessment to culminate experiences that give information about students' knowledge and skills.

Coombe, et al. (2007) classify language assessment into two categories: informal and formal. Among the informal assessments there are classroom (low-stakes), criterion- reference, achievement, direct, subjective, formative, and alternative, authentic whereas standardized (high-stakes), norm-reference, proficiency, indirect, objective, summative, and traditional test are part of the formal assessments category.

This paper will focus on formal classroom assessment instruments, more specifically reading and listening tests. There was an initial decision to work on the tests designed to evaluate

these skills since either the selection of texts or the questions asked in these exams motivated considerable questions among the team of teachers who applied them.

In general, test design should be guided by assessment principles, according to Coombe, et al. (2007) the eight principles that govern test design are: usefulness, validity, reliability, practicality, washback, authenticity, transparency, and security.

Tests are one way to follow up students learning and many decisions are made based on the results students achieve in them. The quality of tests will determine the quality of the conclusions that can be drawn from them once applied. Therefore, "test usefulness provides a kind of metric by which we can evaluate not only the tests that we develop and use, but also all aspects of test development and use" (Bachman, 2001, p. 17). Consequently, usefulness is one of the most important qualities in testing. Furthermore, every language test should have a particular purpose, a specific group to which it will be applied, and a specific language in mind.

The second principle designers of good assessments should take into account is validity. This principle is based on the idea that teachers or instructors should assess what they taught in the same way they taught it. This principle contemplates content, construct, and face validity. Content validity "means that the test assesses the course content and outcomes using formats familiar to the students" (p. 22). Brown (2004) explains this concept saying that a reading test, for example, should test only the students' ability to read, not other abilities such as their vision or previous knowledge of a subject.

To Coombe, et al. (2007) construct validity "refers to the "fit" between the underlying theories and methodology of language learning and the type of assessment" (p. 22). Brown (2004) illustrates this concept with an example of an assessment on oral fluency. He establishes that to possess construct validity this assessment "should account for the various components of

fluency: speed, rhythm, juncture, (lack of) hesitations, and other elements within the construct of fluency" (p. 33).

Face validity means that "the test looks as though it measures what it is supposed to measure" (Coombe, et al. 2007, p. 22). In other words, a test should look like a test and measure what it claims to be measuring; all this aligned with the outcomes and objectives. This judgment is made with the subjective eyes of the test takers and the administrative staff in charge of deciding the use of the assessment (Brown 2004).

The third principle, reliability, is seen from the perspective of consistency of the results obtained by test takers. With this principle, it is assumed as a fact that a test would offer similar results if given at another time in similar conditions. "A common theme in the assessment literature is the idea that reliability and validity are closely interlocked. While reliability focuses on the empirical aspects of the measurement process, validity focuses on the theoretical aspects and interweaves these concepts with the empirical ones, Davies et al., 1999" (Coombe, et al. 2007, p. 3).

Brown (2004) distinguishes four factors that can affect test reliability, namely: the student, the scoring (rater reliability), the test administration, and the test itself. "The most common learner-related issue in reliability is caused by temporary illness, fatigue, a "bad day," anxiety, and other physical or psychological factors, which may make an observed score deviate from one's "true" score" (p. 28). Brown (2004) states that there are some factors that interfere with intra-rater reliability, which are: "Human error, such as lack of adherence to scoring criteria, inexperience, inattention, preconceived biases, subjectivity". Inter-rater reliability "occurs when two or more scorers yield consistent scores of the same test" (p. 28).

Sources of test unreliability can be “photocopying variations, the amount of light in different parts of the room, variations in temperature, and even the condition of desks and chairs” (Brown 2004, p. 29). In classroom assessment, which is the focus of this study, test unreliability can be derived from many different sources, including rater bias. This is a common situation in subjective test with open ended responses, for example, essay responses, in which the teacher has to determine the correct and incorrect answers, whereas objective tests follow a format and have determined responses and this increases the test reliability (Brown, 2004)

Practicality is the fourth principle good assessments should be governed by. “A good classroom test should be "teacher friendly." A teacher should be able to develop, administer, and mark it within the available time and with available resources” (Coombe, et al. 2007, p. 24). After giving a test to students, meaningful feedback is only possible when instructors return promptly the results of the test, and if we are dealing with an impractical test, this will not be possible. Brown (2004) states that practicality “refers to the logistical, down-to-earth, administrative issues involved in making, giving, and scoring an assessment instrument. These include costs, the amount of time it takes to construct and to administer, ease of scoring, and ease of interpreting/reporting the results" (p. 26)

As mentioned previously, the main purpose of assessing students is to get necessary feedback not only of their learning but also of the process done by the teacher. The fifth principle is washback, and it is mainly concerned with the “effect of testing on teaching and learning. Washback is generally said to be positive or negative” (Coombe, et al. 2007, p. 25). Negative washback is seen when students study only for the exam or what they need to know for the exam. The term for this situation is “test-driven curricula.” Positive washback, also called guided washback, “benefits teachers, students, and administrators because it assumes that testing

and curriculum design are both based on clear course outcomes that are known to both students and teachers/testers" (Coombe, et al. 2007, p. 25).

Brown (2004) explains that all classroom-based issues, including informal and formal assessment can derive useful washback, being informal performance assessment more beneficial because the teacher provides interactive feedback. "Formal tests can also have positive washback, but they provide no beneficial washback if the students receive a simple letter grade or a single overall numerical score" (p. 38). Brown goes further and states that teachers have an enormous challenge and is to create classroom assessments that can work as learning devices, in which incorrect responses become opportunities to continue learning and the correct ones need to be praised.

Authenticity is the sixth principle of good assessments. According to Coombe, et al. (2007) a well-designed assessment "strives to use formats and tasks that mirror the types of situations in which students would authentically use the target language. Whenever possible, teachers should attempt to use authentic materials in testing language skills" (p. 25). Brown declares that two or three decades ago, isolated, unconnected and boring items were acceptable in testing but things are different now. Authenticity is a vital pinion in the gear an assessment represents. Bachman and Palmer (1996 in Brown 2004) defined authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a target language task" (p. 23). In their work, an agenda for identifying target language tasks and for transforming them into valid test items is suggested.

Among all the qualities an assessment should have, transparency, the seventh principle, is key to allow students to know relevant information about the test, such as outcomes to be

evaluated, formats used, weighting of items and sections, time allowed to complete the test, and grading criteria (Coombe, et al. 2007).

The last principle acknowledged by Coombe, et al. (2007) is security. This principle plays an important role in other two principles, namely reliability and validity. "If a teacher invests time and energy in developing good tests that accurately reflect the course outcomes, then it is desirable to be able to recycle the test materials" (p. 25). As a consequence, if test designers and administrators want to use more than once a test in which they invested a considerable amount of time and energy, it is necessary to maintain a secure procedure in the whole process of applying the test and giving feedback.

Now that we have seen in detail what assessment is and what it entails, and also its difference with testing and the principles good assessments should have, it is time to revise the term test specifications and their possible benefits and constraints on test design in terms of text selection and quality of questions asked, since this is the objective of this study.

The term test specification is not new. It was probably taken from the "industrial concept of a "specification" for a factory product or engineering objective, the earliest mention we have located in educational and psychological assessment was by Ruch" in 1929 (Davidson, et al. 2001, p.4). Back then the term specification served the same purpose, "provide an efficient generative blueprint by which many similar instances of the same assessment task can be generated" (Davidson, et al. 2001, p. 4).

Fulcher and Davidson (2007) define test specifications (test specs, hereafter) as "generative explanatory documents for the creation of test tasks. Specs tell us the nuts and bolts of how to phrase the test items, how to structure the test layout, how to locate the passages, and how to make a host of difficult choices as we prepare test materials" (p. 52), and the most

important thing is that they serve as a means for us to be aware of the rationale behind the multiple choices we make regarding test design.

According to Fulcher and Davidson (2007), one of the benefits of test specs is test equivalence. To explain test equivalence, we find an example of a situation in which a teacher had a specific test task and wanted an equivalent task, which means same difficulty level, same testing objective, but different content. So, a different test is needed but with the same results. Perhaps test security is an issue, and a new version of the same test is required but with the same assurance of reliability and validity. With this example, we have condensed the original purpose of Specs.

Test Specs are a very important topic in this study since what is being analyzed is the role they play in the selection of language level appropriate texts and the creation quality items. Test Specs have been created for reading and listening assessments, this is why it is important to have some information on how to assess reading and listening. Brown (2004) acknowledges the listening performance as an "invisible, inaudible process of internalizing meaning from the auditory signals being transmitted to the ear and brain" (p. 118). It is invisible and inaudible because teachers can only perceive students' responses to what they listen but not the process itself.

Brown (2004) identifies four types of listening assessment, namely: intensive (listening for perception of the components), responsive (listening to a relatively stretch of language), selective (processing stretches of discourse), and extensive (listening to develop a top-down global understanding of spoken language).

In a like manner, the process of reading cannot be seen, nor can a specific product of reading be observed (Brown, 2004). "The assessment of reading can imply the assessment of a storehouse of reading strategies, aside from simply testing the ultimate achievement of comprehension of a written text" (Brown, 2004, p. 188). The types of reading assessment are: perceptive (attend to the components of larger stretches of discourse), selective (recognition of lexical, grammatical, or discourse features of language), interactive (negotiating meaning), and extensive (texts of more than a page).

The items quality is a relevant topic in this study because, as mentioned before, the role of Test Specs in the creation of quality items is being analyzed. Fulcher and Davidson (2007) argue that communicative language testing should only contain items and tasks that mirror language use in the 'real world', "reflecting the actual purposes of real-world communication, in clearly defined contexts, using input and prompts that had not been adapted for use with second-language speakers" (p. 63). The creation of valid items is not an easy endeavor, in fact, Mullis & Martin (2013) explain that item writing "demands considerable discipline in working within the assessment frameworks and following the guidelines for item construction" (p. 2).

5. Methodology

Having a design or plan is an essential factor in research because it contributes with appropriate orientations that maintain the focus and development of a given inquiry. This research plan or design depends on the kind of methodology researchers use to make decisions that help them to explore the area of study in search for answers to questions. Therefore, there are different methodologies that can be implemented when doing research.

According to Herbert (as cited in Seliger & Shohamy, 1989), those methodologies differ not only in the connection that researchers and the study have, the way data is collected, or the context in which they are immersed, but also how they can define certain answers. Thus, while some methodologies are very explicit, others can be quite general, or when some of them are lightly manipulated, others do not possess any type of intervention. Likewise, Herbert (2001) in Seliger & Shohamy (1989), presents *qualitative, quantitative, experimental*, and *descriptive* research as methodologies to efficiently carry out a research project.

He defines qualitative research as the study of individuals' performance where no definite answers or generalizations exist but great approximations of reality. Therefore, questionable information is presented in this type of research since knowledge is treated with humility, and it is clear that what works for a particular context does not necessarily work for another.

On the other hand, Shulman (1980) states that "when our interest is in the normative acquisition behavior of a population, quantification represents a reality for that group. Such a reality may be generalizable to other groups, assuming that sampling procedures are adequate" (p. 115). Therefore, since every detail has to be carefully measured, quantitative sometimes appears to be more complete than qualitative research especially when making decisions and

presenting gathered information. Quantification other times complements qualitative research at the moment of categorizing data collected.

Additionally, as qualitative, descriptive research describes a natural phenomenon without any type of intrusion or manipulation. However, both types of research implement different views and approaches to carry out a project. Kamil et al (1985) mention that one of the differences between descriptive research and qualitative research is the way in which they collect and analyze data. For instance, in qualitative research, research question and data collection are designed before the research starts. Thus, a hypothesis is proposed to make proper decisions. Unlike qualitative, descriptive research not only uses existing data but also possesses preconceived hypothesis and questions that go from general to specific focus.

In contrast to those, experimental research is totally different because it manipulates and controls significant measures that help validate results. In this type of research, the role of the researchers does affect the behavior of the subject and answers might be easily expected.

Furthermore, Teddlie & Tashakkori (2003) discuss a new concept called *mixed research* also known as the *Third Methodological Movement* because it follows the principles of both qualitative and quantitative researches. In other words, mixed research involves the use of both types of research in a single study. Additionally, Johnson and Turner (2003) say that the main purpose of Mixed Research is to combine qualitative and quantitative approaches, methods and strategies with complementary strengths and no overlapping weaknesses. This means that Mixed Research takes the best of each type of research to assertively implement diverse approaches and methods that can be performed in multiple and innovative manners.

If analyzed previous contributions about Mixed Research, it is worth mentioning that other authors agree on this concept. According to Campbell & Fiske (1959), Mixed Research

integrates both qualitative and quantitative studies but not necessarily at the same time. In fact, decisions are made based on sequence of data collection, relative priorities of the study, and how certain quantitative or qualitative components take place. Likewise, since researchers possess more flexibility to make decisions about the type of research to use, they can also focus on more specific aspects such as a particular language structure or a particular language behavior, and these patterns can allow researchers efficiently narrow the research study at the proper time.

According to Creswell et al (2007), there are four mixed methods designs. They are Triangulation, Embedded, Explanatory and Exploratory designs. The method that suits the best in this type of research is Exploratory mixed method because it starts with Qualitative approach, and discoveries found are explained and validated by using Quantitative methods. This design usually implements a quite variety of standardized Qualitative and Quantitative instruments.

Since the purpose of this paper is to explore the role that the use of test specifications plays on test design in terms of the selection of level appropriate texts and the creation of level appropriate questions in an EFL program, mixed research is considered as the method to apply due to the fact that both qualitative and quantitative approaches are integrated, and it also explains qualitative aspects in a quantitative way.

This research is conducted with *reading* and *listening* formal summative assessments for levels two and four in a language program at a private University in Colombia. The methodology used allowed for the implementation of data collection instruments and its analysis in order to obtain relevant information. Therefore, a checklist was used as the main data collection instrument for this research.

It is worth mentioning that a document revision of the language program was done and used as a reference for the design and implementation of the checklist. Document revision is a

vital part of this research because mute evidence (written texts) can be analyzed and interpreted without the influence of deliberate dialogues and comments as it might happen in interviews and surveys. According to Denzin & Lincoln (1994), in document revision "there is often no possibility of interaction with spoken emic "insider" as opposed to etic "outsider" perspectives" (p703). Thus, there is no chance of alteration and interventions in document analysis since unspoken words are revised and isolated of subjectivity; the researcher has a criteria of revision to neutrally focus on certain written aspects and not have contact with verbal effects.

The document revision implemented in this research was conducted to study all the parameters of the language program that the university aims, especially in levels II and IV, such as the syllabus and the outcomes of the levels, the Assessment Handbook and the Test Specs Guidelines. This revision was done to create the checklist and compare and contrast the tests that were designed with the use of Assessment Handbook with the tests designed after the implementation of Test Specs. In other words, the Assessment Handbook, the Tests Specs, the syllabus and outcomes of levels II and IV were revised with the purpose of exploring the role that the use of test specifications plays on test design in terms of the selection of level appropriate texts and the creation of level appropriate questions in the EFL program of the University.

The Assessment Handbook was the official guideline to design tests before the second semester of 2016 at the University. However, the creation of Tests Specs was required since the Assessment Handbook was quite general. These Test Specs are more precise in terms of assessment principles such as validity, reliability, practicality, and so forth. The purpose of the paper is also concerned with validity, and Test Specs are likewise in function of the same principle.

In the same way, based on the document revision previously mentioned, a checklist was designed as the main instrument to apply in this research. The checklist more specifically aimed to analyze reading and listening tests of level II and IV. This instrument evaluates the reading and listening tests that were designed with the use of Assessment Handbook (201510-201530 and 201610) and the reading and listening tests designed after the implementation of Test Specs (201630 and 201710) of both levels II and IV. This checklist revises all tests in terms of validity, text language appropriacy, and quality of test items. It has a set of eighteen Yes/No questions that validate the role that Test Specs plays on test design in terms of the selection of level appropriate texts and the creation of level appropriate questions in the EFL program at the University. The answers to these questions in the checklist were categorized under "Yes" if the category was identified in the tests analyzed, "No" if they were not, and "NA" if the category of the checklist did not apply.

Similarly, after filling each checklist of level II and IV, "Yes", "No", and "N/A" answers were tabulated per question in all the exams prior and after Test Specs. These questions were also divided into three categories (Validity, Text language level appropriacy and Test items quality). Thus, graphs were created per question of each category, and they show the results (Yes, No, NA) of each exam prior and after Test Specs. This process was developed with listening and reading tests of level II and IV in the language program of the university.

Additionally, the tables presented further on show the analysis made of the tests designed before and after the implementation of the Test Specs. They are presented by skill, by level, and by the categories mentioned above. The tables do not only show the amount of times the tests complied or not but they also illustrate the percentages of the times tests met the criteria defined "Yes" or whether they did not "No." These percentages were taken from the relation of the

number of questions in each category with the number of times these questions were answered "Yes" and "No."

Formula sample:

- **Level II Listening Test (201710) Validity (4 questions)**

201710		
NA	***No***	***Yes***
0	**1**	**3**
0%	**25%**	**75%**

4 Questions = 100%
1 **No** answer =?

$$\frac{1 * 100}{4} = \frac{100}{4} = \mathbf{25\%}$$

- **Level II Listening Test (201710) Text Language Level Appropriacy (5 questions)**

201710		
NA	***No***	***Yes***
0	**1**	**4**
0%	**20%**	**80%**

5 Questions = 100%

4 **Yes** answer =?

$$\frac{4 * 100}{5} = \frac{400}{5} = 80\mathbf{\%}$$

- **Level II Listening Test (201710) Test Items Quality (9 questions)**

201710		
NA	***No***	***Yes***
1	**1**	**7**
11.1%	**11.1%**	**77.7%**

9 Questions = 100%

7 Yes answer =?

$$\frac{7*100}{9} = \frac{700}{9} = \mathbf{77.7\%}$$

In other words, documents such as the syllabus, the outcomes of the levels, the Assessment Handbook, the Test Specs and the reading and listening tests of level II and IV were taken into account to answer the questions formulated in the checklist. After analyzing all the tests and documents with the implementation of the checklist, common findings were established in tables, and conclusions were written down as a manner of systematizing the analysis, and illustrating the impact Test Specs have in the design of these tests.

The purpose of validity questions in the checklist was to revise how valid reading and listening tests of level II and IV were if they measured what they intended to measure, had reasonable number of questions, well distributed items and followed specific guidelines.

I. VALIDITY	YES	NO	NA	COMMENTS
1. Does the test accurately measure what it intends to measure?				
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?				
3. Are the items well distributed to make the test valid?				
4. Does test construction follow specific guidelines?				

The first question of validity asks whether the test accurately measures what it intends to measure. To be more precise, the checklist revises if the test is actually evaluating the estimated outcomes and topics of level II and IV in listening and reading skills. The second question of validity checks if the test has reasonable number of questions that can be completed by students within the expected time frame. This means that the test should ask reasonable number of questions taking into account the amount of time given to students. The third question reviews if items are well distributed to make the test valid. In other words, items and sections of the test should all be of similar length and structure. The fourth question asks if test construction follows specific guidelines such as the Assessment Handbook and Test Specs.

Text language level appropriacy questions aim the selection of level appropriate texts on test design. They study if the language in the test is representative of real-world language use, if the items are level appropriate, contextualized and unambiguous, and if the tests were designed with a specific purpose, a particular group of test-takers, and specific language use in mind.

II. TEXTS LANGUAGE LEVEL APPROPRIACY	YES	NO	NA	COMMENTS
1. Is the language in the test representative of real-world language use?				
2. Are the items level appropriate?				
3. Does the test have items that are contextualized rather than isolated?				
4. Does the test contain items/tasks that are unambiguous to the test-taker?				
5. Is the test developed with a specific purpose, a particular group of test-takers, and specific language use in mind?				

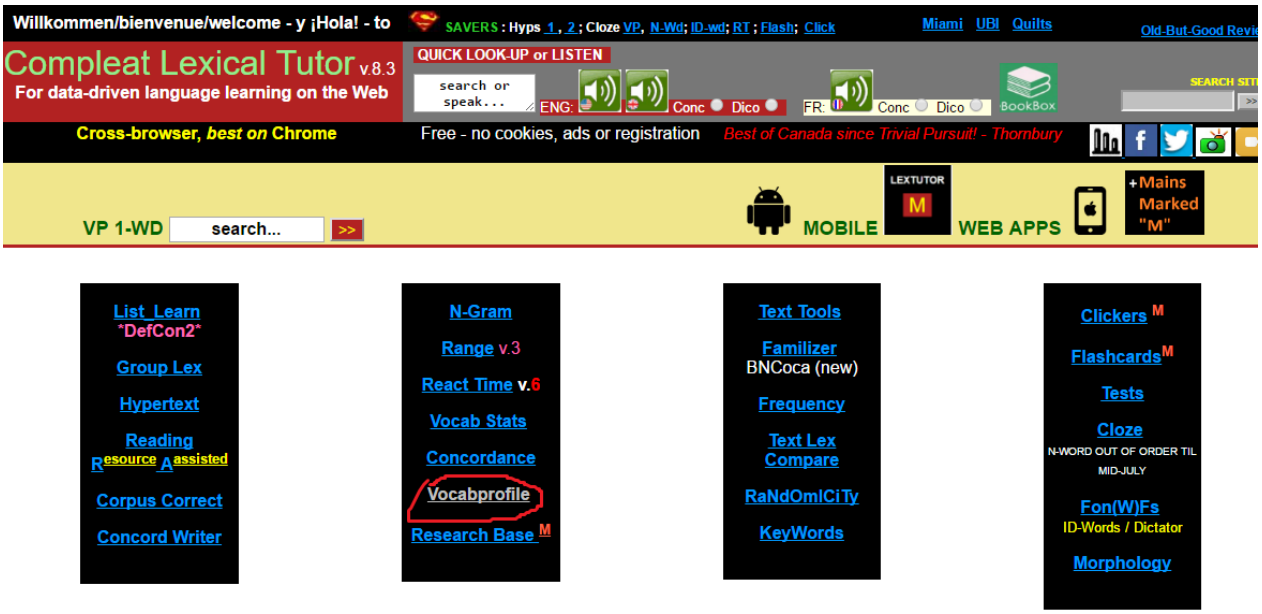
The first question asks whether the language in the test is representative of real-world language use. This means that tests, regardless the listening and reading skills, should reflect the use of actual language that students can face in real life situations.

The second question in the checklist revises if the items of the test are level appropriate. To be more precise, texts need to fit the required level evaluated in the exams (Level II and Level IV). A Website was used to check the lexical level appropriateness of listening and reading tests of both levels II and IV at a lexical level (http://www.lextutor.ca/). VocabProfilers is the tool used in the website that revises the level appropriacy of tests by breaking texts down into word frequencies in the language. There are three ways to analyze texts in Vocabprofilers: VP-Classic, VP-Kids and VP- Compleat. VP-Compleat is the option chosen in this analysis since it divides text words into frequency levels (K-1 Words, K-2 Words, K-3 Words, K-4 Words, etc.). For instance, a level II text is expected to have most lexical items classified within the K-1 or K-2 words and for level IV up to K3. Whatever words classify as off-list or academic word list, these are expected to have been explicitly taught in class or they might be cognates easily identified by students; if they have predominant results at the frequency band expected level, texts can be said to be appropriate for the corresponding level.

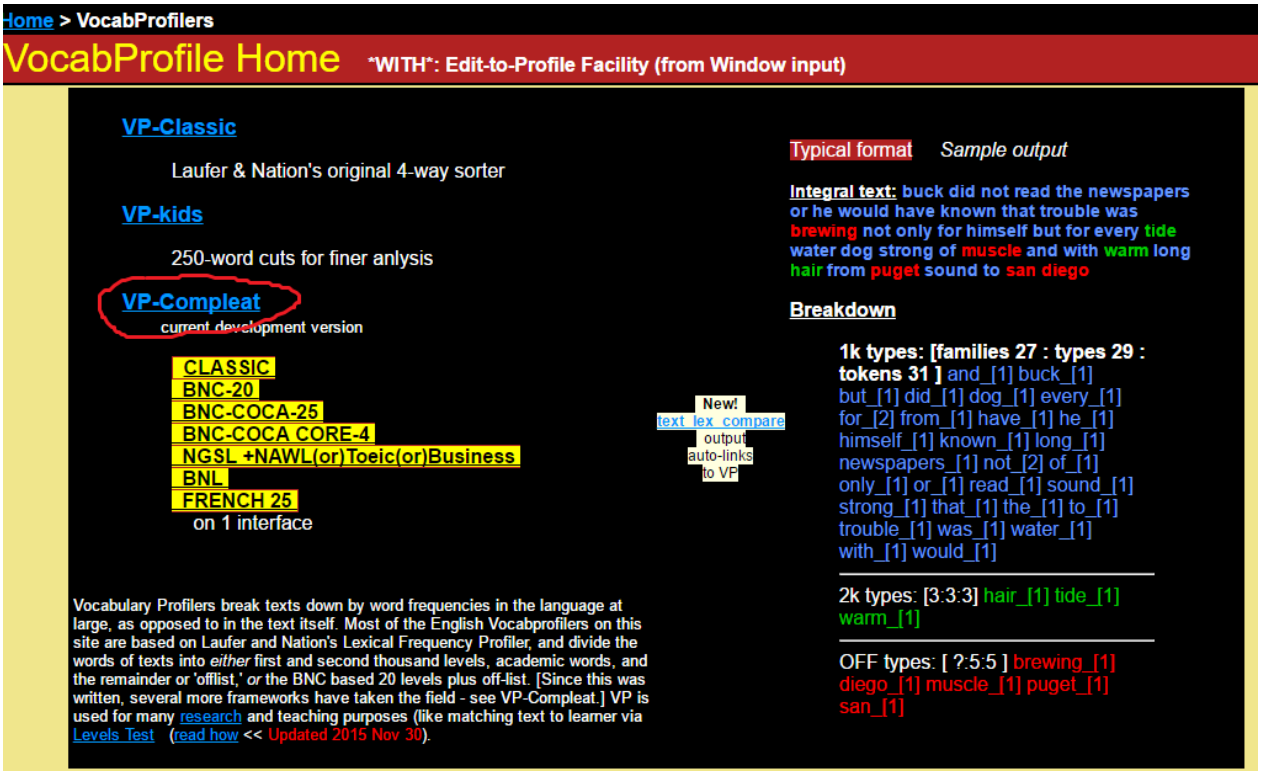
An example of the steps followed to analyze texts is given below:

- **Step 1:** Go to the link http://www.lextutor.ca/

Step 2: Click on Vocabprofile

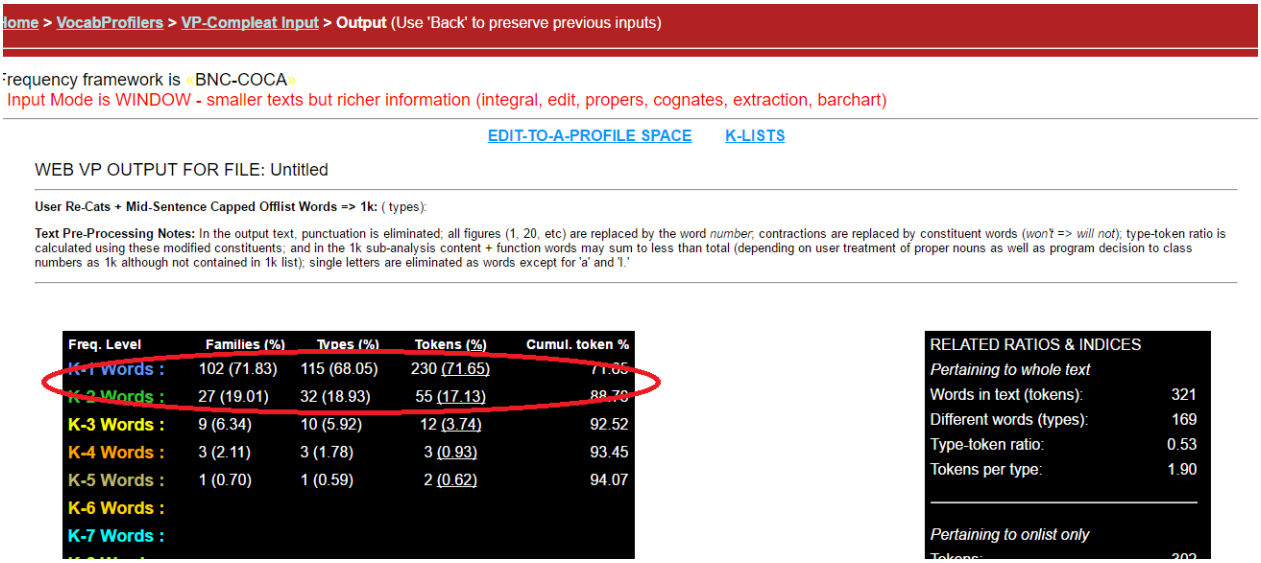


- **Step 3:** Once you are in Vocabprofile, click on VP-Compleat



- **Step 4:** Enter the text selected and click on **submit _window**

- **Step 5:** Level II reading test 201510 sample results.



ome > VocabProfilers > VP-Compleat Input > **Output** (Use 'Back' to preserve previous inputs)

requency framework is «BNC-COCA»
Input Mode is WINDOW - smaller texts but richer information (integral, edit, propers, cognates, extraction, barchart)

EDIT-TO-A-PROFILE SPACE K-LISTS

WEB VP OUTPUT FOR FILE: Untitled

User Re-Cats + Mid-Sentence Capped Offlist Words => 1k: (types):

Text Pre-Processing Notes: In the output text, punctuation is eliminated; all figures (1, 20, etc) are replaced by the word *number*; contractions are replaced by constituent words (*won't => will not*); type-token ratio is calculated using these modified constituents; and in the 1k sub-analysis content + function words may sum to less than total (depending on user treatment of proper nouns as well as program decision to class numbers as 1k although not contained in 1k list); single letters are eliminated as words except for 'a' and 'I.'

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	102 (71.83)	115 (68.05)	230 (71.65)	71.65
K-2 Words :	27 (19.01)	32 (18.93)	55 (17.13)	88.79
K-3 Words :	9 (6.34)	10 (5.92)	12 (3.74)	92.52
K-4 Words :	3 (2.11)	3 (1.78)	3 (0.93)	93.45
K-5 Words :	1 (0.70)	1 (0.59)	2 (0.62)	94.07
K-6 Words :				
K-7 Words :				

RELATED RATIOS & INDICES

Pertaining to whole text

Words in text (tokens):	321
Different words (types):	169
Type-token ratio:	0.53
Tokens per type:	1.90

Pertaining to onlist only

As the sample shows, the most predominant results are placed in K-1 and K-2 words which means that the text selected is lexically suitable for reading test of level II.

The third question checks if tests have items that are contextualized rather than isolated. In other words, items should be contextualized with the listening or reading texts and the outcomes of the level II and IV. The fourth question also revises whether the test contains items/tasks that are unambiguous to the test-taker. That means that tests do not provide items or tasks that are unclear and imprecise to students. The last question of text language level appropriacy is if the test is developed with a specific purpose, a particular group of test-takers, and specific language use in mind. Likewise, the outcomes of each level, the conditions of the students taking the tests and the topics of the books are reflected in the listening and reading texts chosen and the items included.

Test items quality questions validate the creation of level appropriate questions on test design in the EFL program of the University. Aspects as the stems, options, text dependent questions, distractors, matching exercises, and parallelism are revised here.

III. TEST ITEMS QUALITY	YES	NO	NA	COMMENTS
1. Is the stem clear and precise (it clearly indicates the kinds of answer students need to give)?				
2. Is each option clearly identified as the answer to the question asked?				
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)				
4. Is the answer to each question text dependent (it does not depend on other stems and keys)				
5. Are the options of the questions parallel? (formatting)				
6. Are the questions formulated positively?				
7. Are distractors well designed?				
8. Are the numbers of questions in numerical order?				
9. Do matching exercises have two extra options?				

The first question of test items quality asks whether the stem is clear and precise. This means that stems clearly indicate the kinds of answers students need to give. The second question revises if each option is clearly identified as the answer to the question asked. In other words, tests should present clear options that can be distinguished as the answers to the questions in the tests. The third question checks if the answer to each question is text dependent which means that the answers do not depend on students' prior knowledge in order to be solved. The fourth question also checks if the answer to each question is text dependent in terms of the dependency they can have on other stems and keys.

The fifth question asks whether the options of the questions are parallel. To be more exact, questions should start with the same parts of the speech (nouns, verbs, adjectives) and they should follow the same structure. The sixth question asks if the questions are formulated positively. Formulating affirmative statements is important to avoid confusing students because negative statements might lead to students' confusion if these have not been properly practiced in class. The seventh question revises if distractors are well designed. Distractors are very important in multiple choice questions because they do not only challenge students to identify what the correct answer is, but they also foster critical analysis in the application of the exams.

The eighth question checks if the numbers of questions are in numerical order. Numerical order is important to avoid confusing students because it gives consistency and makes tests easier to follow. And the last question revises if matching exercises have two extra options. Having extra options in matching exercises is important to prevent students answering items by process of elimination or it helps to avoid double penalization when they make a mistake. In this way, the use of the checklist as the main data collection instrument is crucial because the data resulting from it can inform test specs design and implementation processes, addressing possible strengths or potential problems.

6. Results

This section is aimed to present the analysis made of the tests designed before and after the implementation of the Test Specs. It will follow the order established by the checklist (See Appendix C) and will approach the characteristics of the listening tests from their Validity, Text Language Level Appropriacy and Test Items Quality. The checklist was used to identify whether the tests met the criteria defined "Yes" or whether they did not "No." The graphs presented in this section show the results obtained after analyzing each test with the checklist. The results are presented by skill (listening and reading) and by the categories in which the checklist is divided.

The Test Specs were designed and piloted in Levels II and IV of the EFL Program at a private university in Colombia. This is why the eighteen tests analyzed are part of the assessment process of these two levels. Level II covers the A2.2 CEFR level (High Beginner) and Level IV covers the B1.2 CEFR level (Pre-intermediate). Taking into account the objective of this study, which is analyzing the text language level and items quality prior and after the implementation of Test Specs, the tests chosen were the ones applied in the following terms:

Before Test Specs

Level II

- 201510 (first term)
- 201530 (second term)
- 201610 (first term)

Level IV

- 201530 (second term)
- 201610 (first term)

After Test Specs

Level II

- 201630 (second term)
- 201710 (first term)

Level IV

- 201630 (second term)
- 201710 (first term)

Listening Tests Levels II and IV

This analysis was done after analyzing each of the exams (Level II: 201510, 201530, 201610, 201630 and 201710; level IV: 201530, 201610, 201630 and 201710) with the assistance of the checklist. "Yes", "No", and "N/A" were tabulated per question in all the exams prior and after Test Specs. These questions were also divided into three categories (Validity, Text language level appropriacy and Test items quality). The impact of Test Specs was initially measured by illustrating a graph per question where "Yes", "No", and "N/A" were visibly identified in the listening tests of level II and IV mentioned above. Then, the same process was developed with reading tests of level II and IV previously mentioned.

Validity

The following graphs show the analysis of tests related to validity. Five questions will guide the comparison of exams designed prior and after Test Specs initial implementation.

1. Does the test accurately measure what it intends to measure?

Before Specs

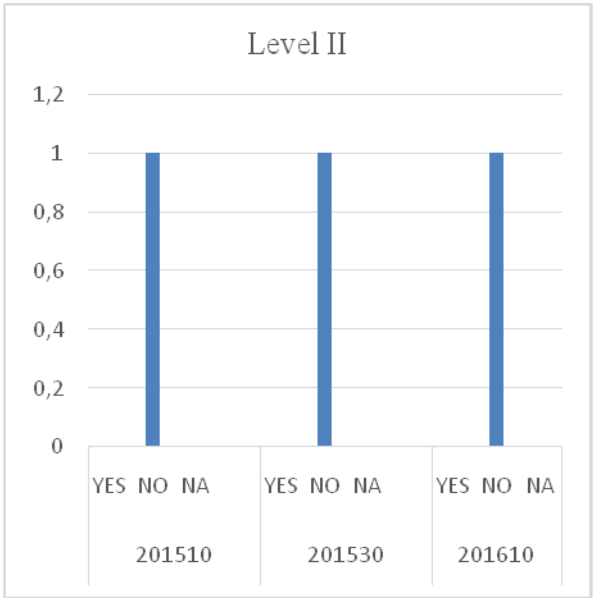


Figure 1

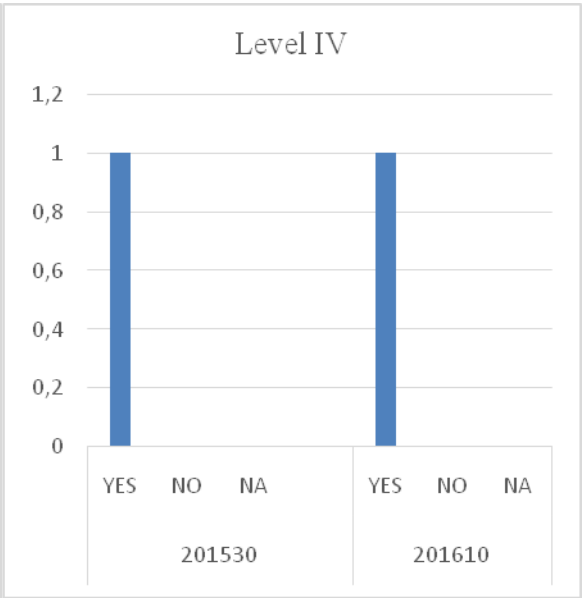


Figure 2

Figures 1 and 2 show whether tests, before test specs, accurately measure what they intend to measure or not. As observed in figure 1, the *201510*, *201530*, and *201610* tests did not measure what they intended to measure, while in figure 2, *201530* and *201610* tests did. This means that level II exams, analyzed for this study, partially failed to comply with the principle of validity when it comes to accurately measure what they intend to measure. Level IV tests complied with this.

After Specs

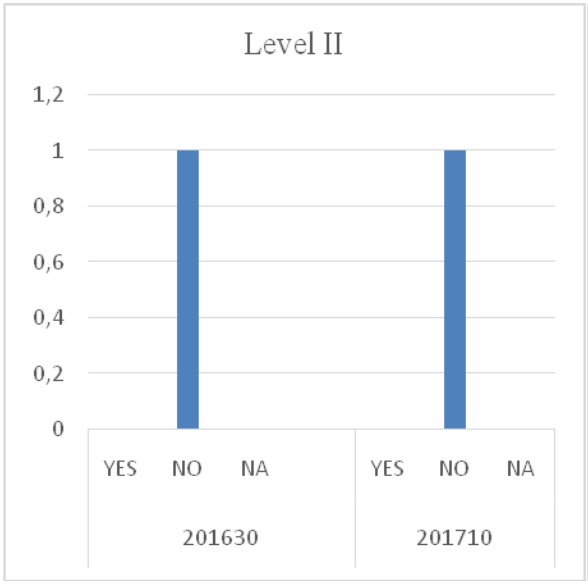


Figure 3

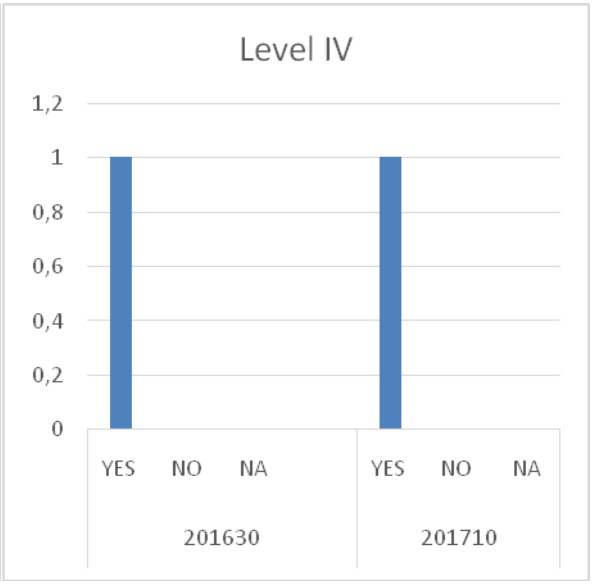


Figure 4

Figures 3 and 4 show if tests, after Test Specs, accurately measure what they intend to measure. As can be seen, none of listening tests of level II *(201630, 201710)* evaluated the estimated outcomes and topics. On the contrary, all of the listening tests of level IV *(201630, 201710*) did measure what they were supposed to measure. In other words, after the implementation of tests specs, Level II listening exams, analyzed for this study, failed to comply with the principle of validity while Level IV listening tests complied with this.

2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?

Before Specs

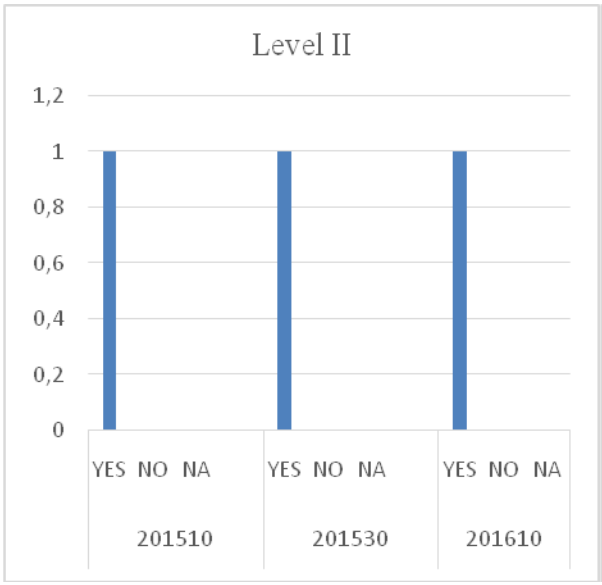


Figure 5

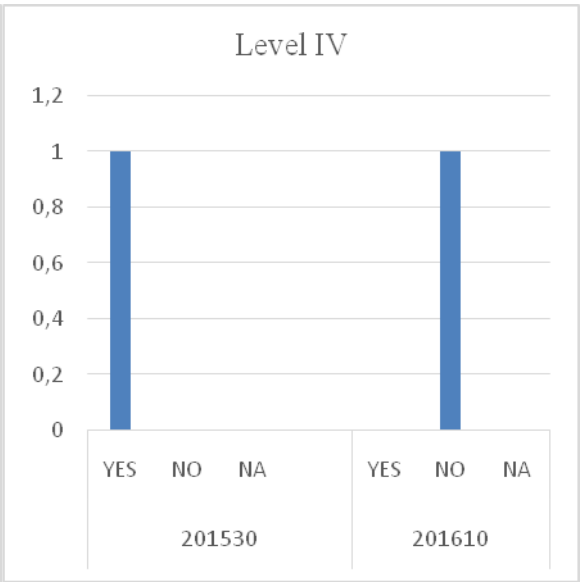


Figure 6

Figures 5 and 6 illustrate the reasonability of the number of questions asked in the tests considering a time frame. It was clear that the *201510*, *201530*, and *201610* tests had a reasonable number of questions in level II. In a like manner, the number of questions of *201530* test of level IV was reasonable whereas in the *201610* test (same level) the number of questions was not reasonable to be completed within the expected time frame. As it can be observed, before the Test Specs, most of the tests asked reasonable number of questions taking into account the amount of time given to students. Even though Assessment Handbook established a fifty minute time frame for students to answer, these guidelines did not specify the approximate number of questions tests should have.

After Specs

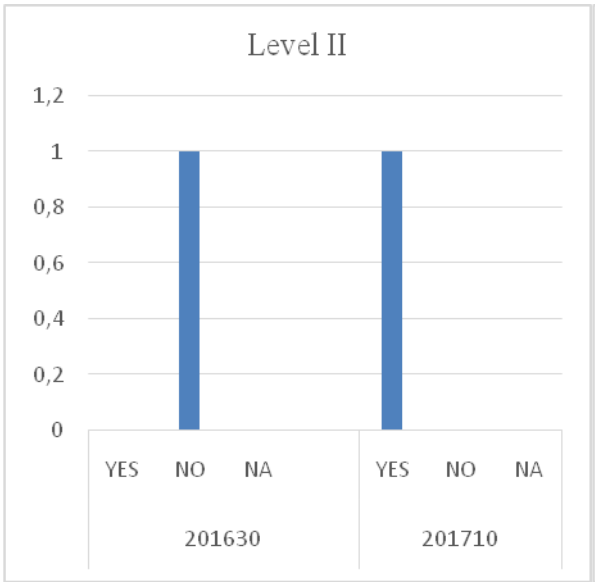


Figure 7

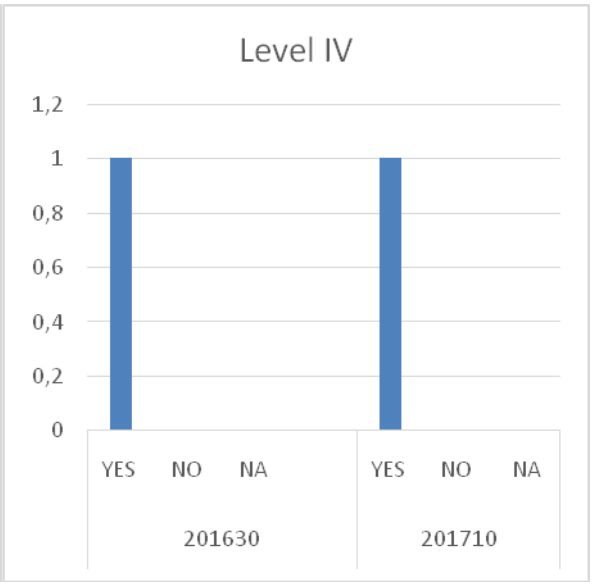


Figure 8

Figures 7 and 8 illustrate the reasonability of the number of questions asked in the tests considering a time frame. As observed in the figures, one of the listening tests of level II *(201630)* did not have reasonable number of questions. On the other hand, the other listening test of level II as well as all listening tests of level IV, after test specs, had reasonable amount of questions for students to complete within the expected time frame. In this manner, most of the tests, after tests specs, asked reasonable number of questions taking into account the amount of time given to students. Even though Test Specs established a fifty-minute time frame for students to answer, these guidelines did not specify the approximate number of questions tests should have.

3. Are the items well distributed to make the test valid? (each section has similar amount of questions and points)

Before Specs

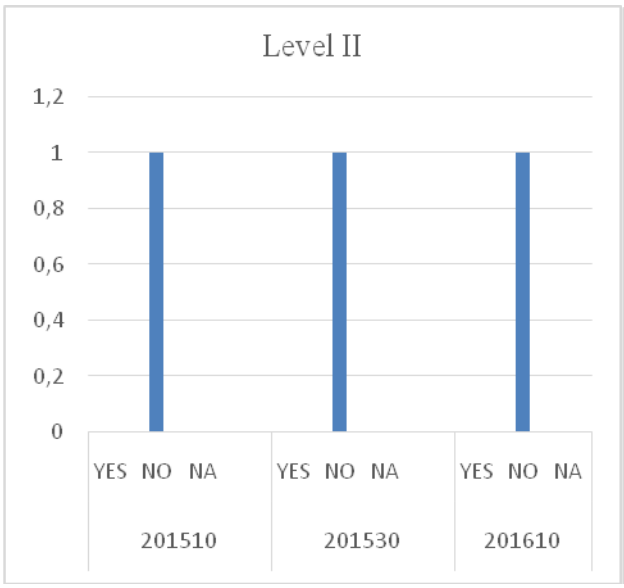


Figure 9

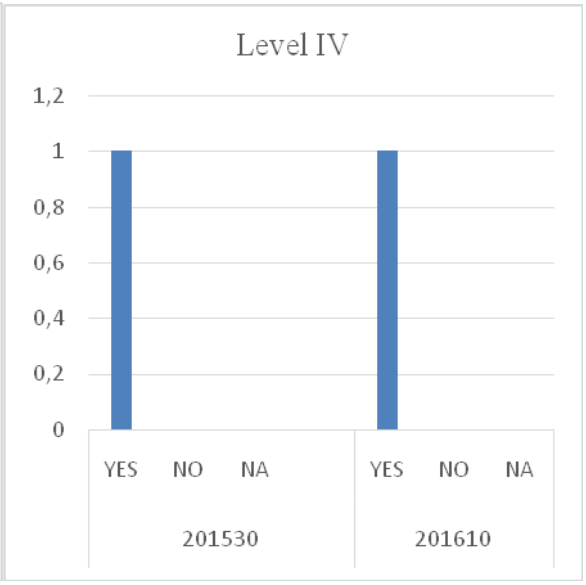
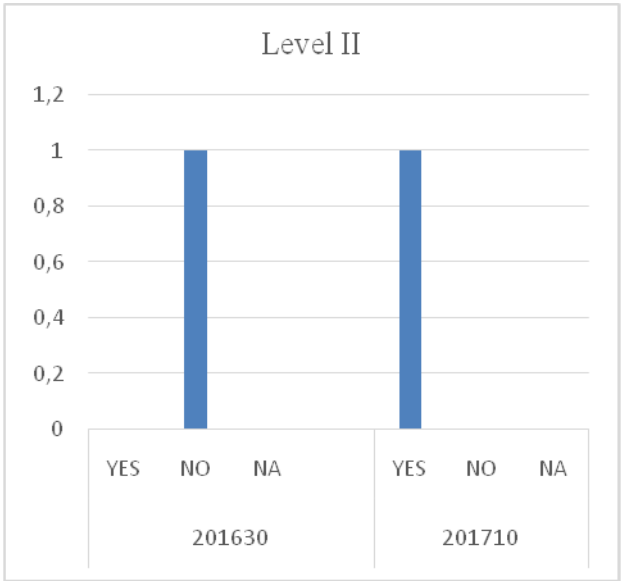


Figure 10

Figures 9 and 10 present the distribution of the items in tests to make them valid. As shown, in the *201510, 201530,* and *201610* tests of level II, the items were not well distributed while in the *201530* and *201610* tests of level IV the items were distributed in a manner that makes the tests valid, before Test Specs. Thus, there is a remarkable difference in the distribution of the questions in tests regarding the levels. According to the guidelines used to design tests, sections should all be of similar length and structure.

After Specs



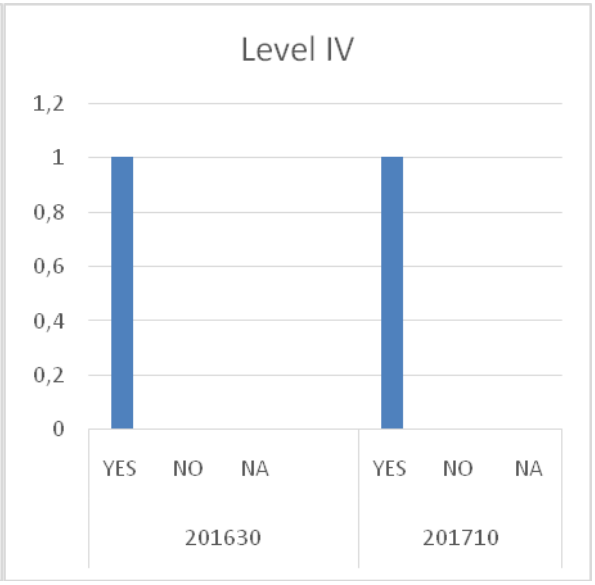


Figure 11 Figure 12

Figures 11 and 12 present the distribution of the items in tests to make them valid (each section has similar amount of questions and points). As shown, in the *201630* test of level II, the items were not well distributed whereas in 201710 test of level II and *201630 and 201710* tests of level IV the items were distributed in a manner that makes the tests valid, after Test Specs. Likewise, most of the tests after test specs distributed items properly. According to the Test Specs used to design tests, sections should all be of similar length and structure.

4. Does test construction follow specific guidelines?

Before Specs

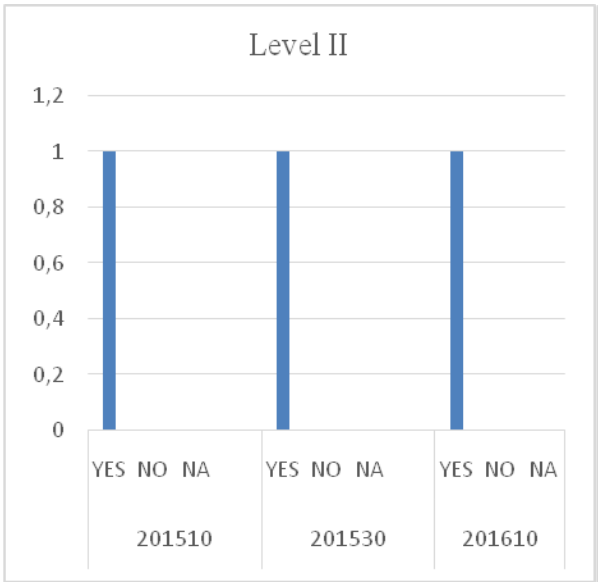


Figure 13

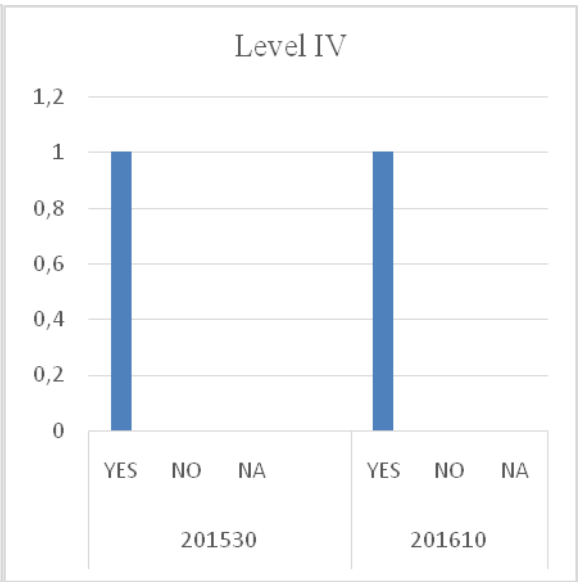
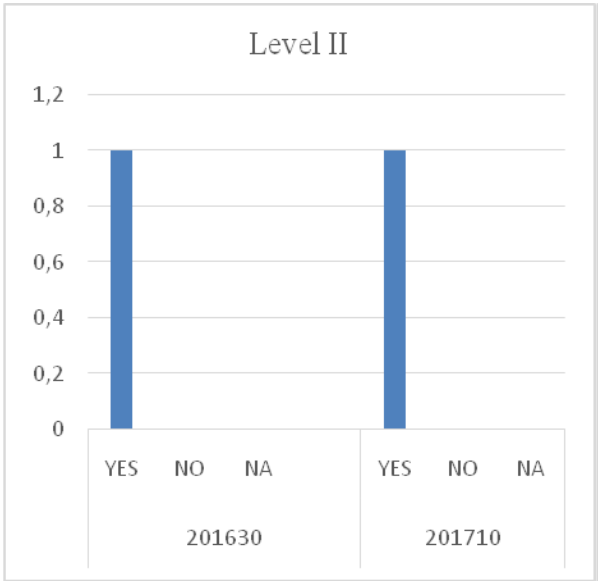


Figure 14

Figures 13 and 14 read whether the design of tests followed specific guidelines or not. As presented in the figures, *201510*, *201530*, and *201610* tests of level II and *201530* and *201610* tests of level IV followed a specific guideline to construct them. So, according to the analysis done for this study, all the tests were designed following detailed instructions (Assessment Handbook).

After Specs



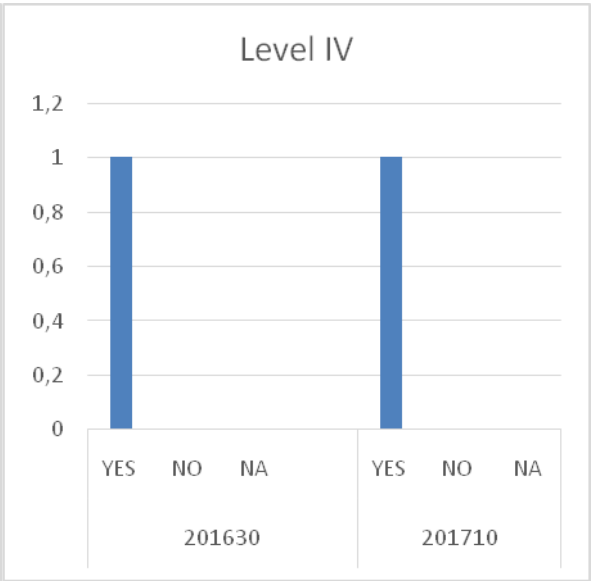


Figure 15 Figure 16

Figures 15 and 16 read whether the design of tests followed specific guidelines or not. As presented in the figures, all tests of level II and level IV *(201630 / 201710)* followed a specific guideline to construct them called Test Specs. Thus, all tests were designed following detailed instructions. (Test Specs)

Texts Language Level Appropriacy

The following graphs show the analysis of the language level appropriacy of the exams designed before and after Test Specs use.

1. Is the language in the test representative of real-world language use?

Before Specs

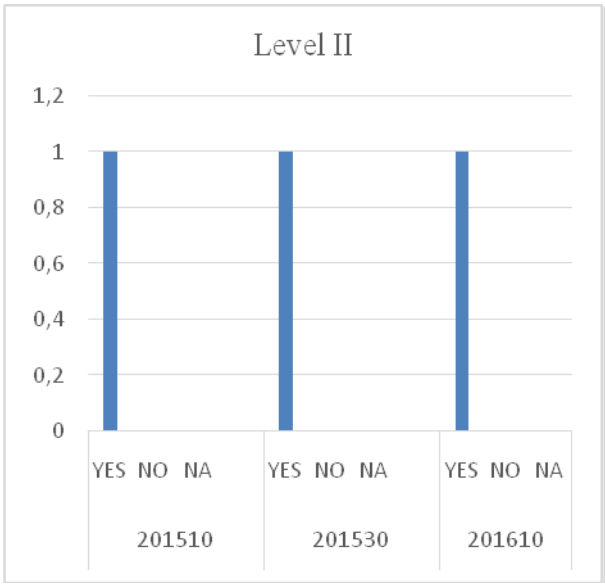


Figure 17

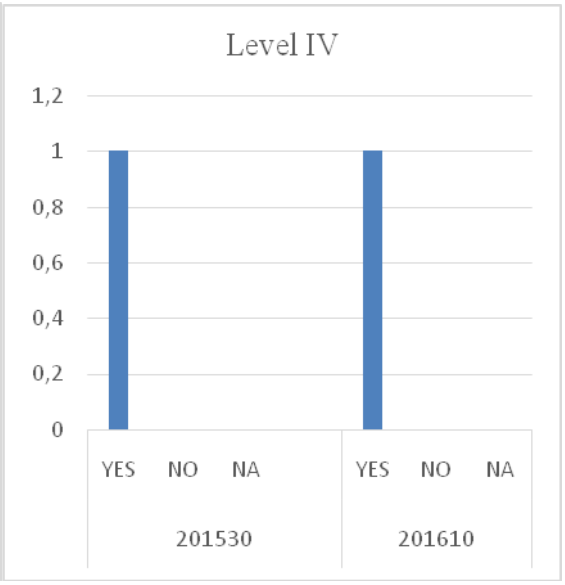


Figure 18

Figures 17 and 18 contain information concerning the language used in tests. As stated in the figures, *201510*, *201530*, and *201610* tests of level II and *201530* and *201610* tests of level IV make use of language representative of real-world language use. Therefore, all the tests examined for this study use language that simulates real-world language. Assessment Handbook states that language should reflect actual use. Some topics are: *Doctor's appointment, Prenuptial agreement and Getting a divorce.*

After Specs

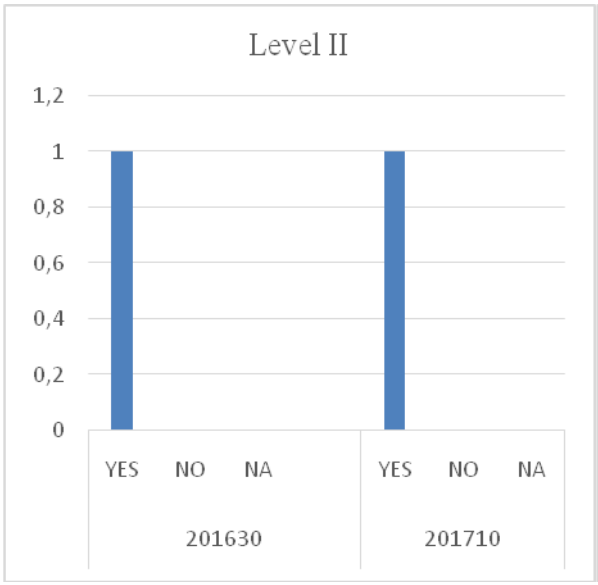


Figure 19

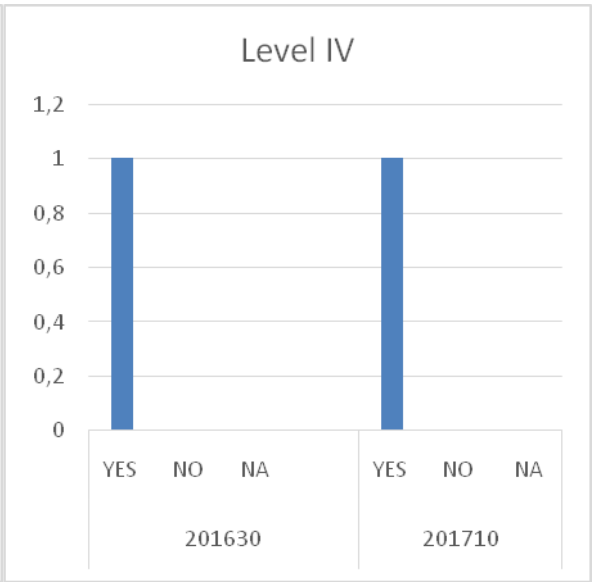


Figure 20

Figures 18 and 19 contain information concerning the language used in tests. As stated in the figures, *201630 and 201710* tests of level II and *201630 and 201710* of level IV make use of language representative of real-world language use. Therefore, all the tests examined for this study use language that simulates the use in real-world. Test Specs state that language should reflect actual use. Some topics were: *The global change effect, Break up, and Recycling*.

2. Are the listening texts level appropriate?

Before Specs

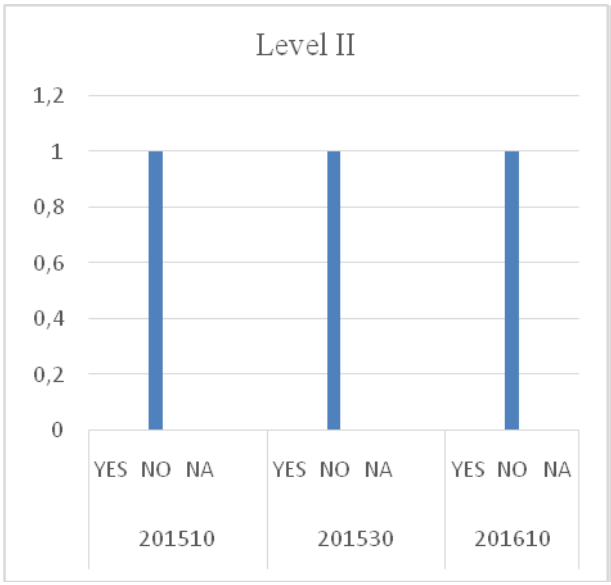


Figure 21

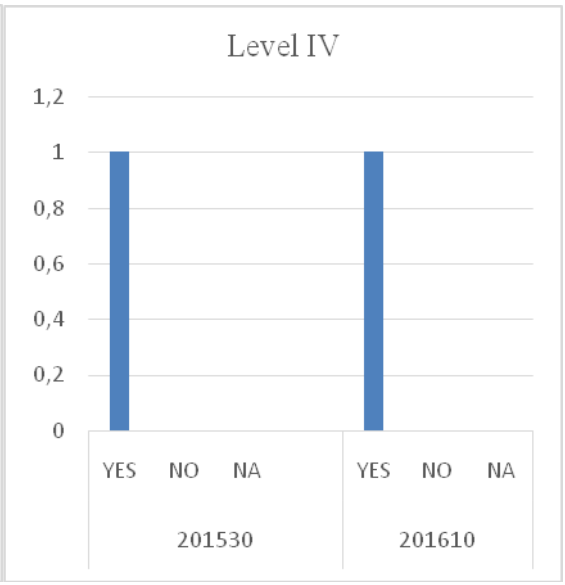


Figure 22

Figures 21 and 22 indicate the level appropriateness of the listening texts used in tests. It can be observed that in *201510*, *201530*, and *201610* tests of level II the listening texts are not appropriate for the level and *201530* and *201610* tests of level IV the listening texts are level appropriate. In other words, the listening texts chosen for level II exams analyzed in this study are not suitable for the level, whereas the listening texts used in the level IV exams fit the required level.

After Specs

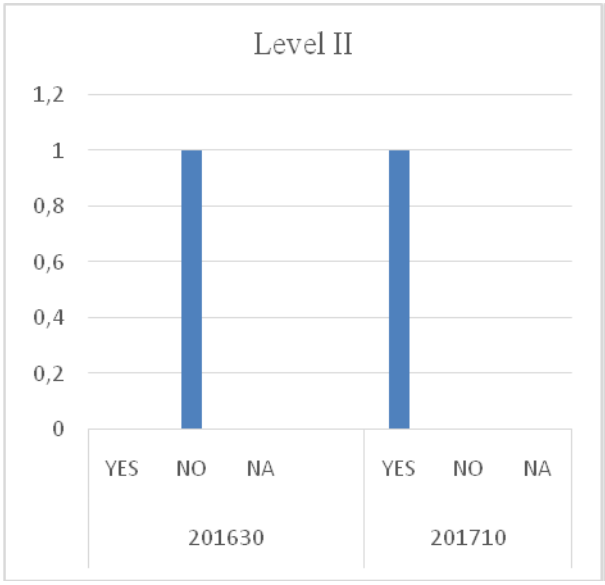


Figure 23

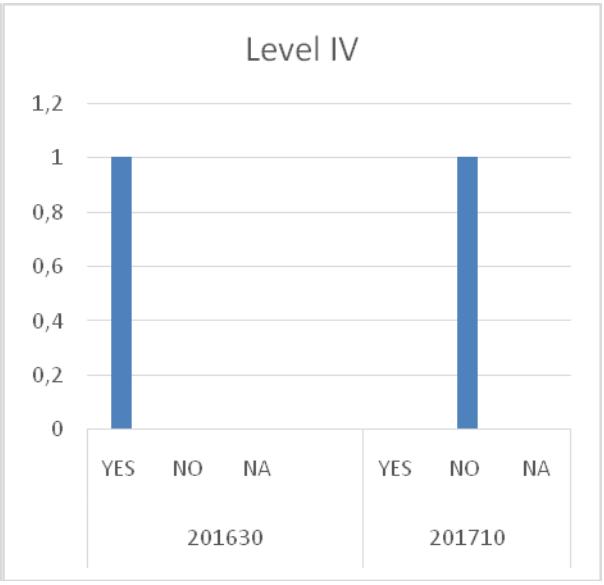


Figure 24

Figures 23 and 24 indicate the level appropriateness of the listening texts used in tests. It can be seen that 201630 test of level II and 201710 test of level IV did not have appropriate texts for the level while 201710 test of level II and 201630 tests of level IV did. To be more precise, two out of the four tests are not suitable for the levels while the other two tests fit the required level.

3. Does the test have items that are contextualized rather than isolated?

Before Specs

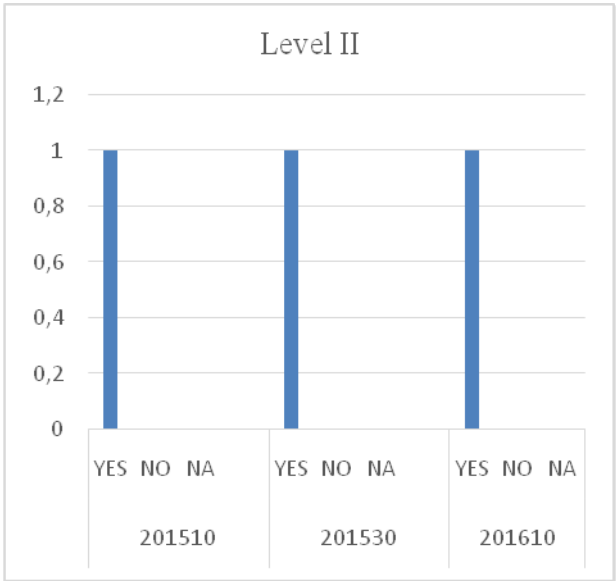


Figure 25

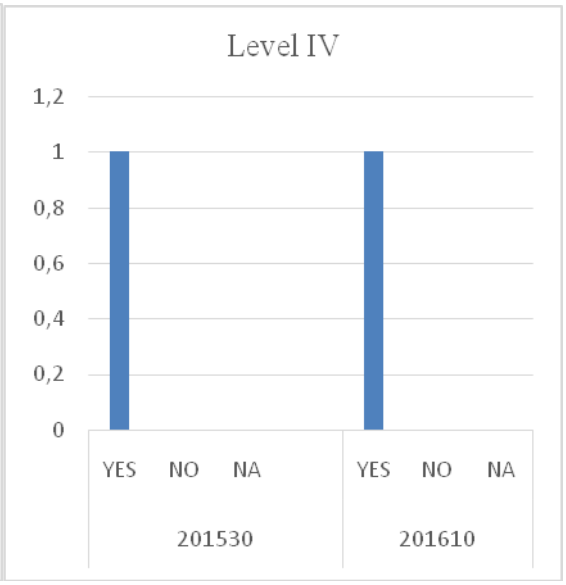


Figure 26

Figures 25 and 26 display if items in the tests are contextualized rather than isolated. In other words, in *201510*, *201530*, and *201610* tests of level II and *201530* and *201610* tests of level IV all the items are contextualized with the listening texts and, at the same time, these are contextualized with the outcomes of the levels. Hence, all the tests analyzed for this project are coherent with the objectives of the levels.

After Specs

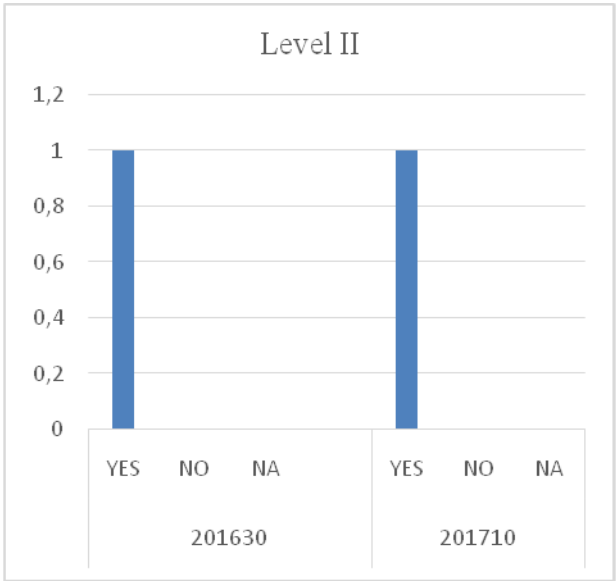


Figure 27

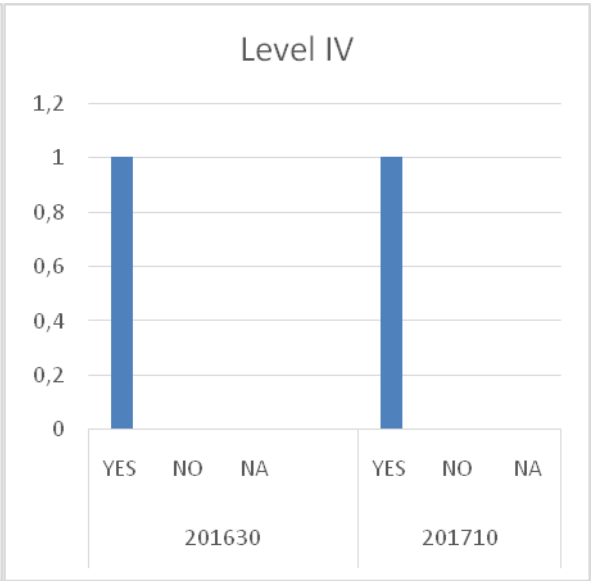


Figure 28

Figures 27 and 28 display if items in the tests are contextualized rather than isolated after the implementation of test specs. Evidently, all *201630 and 201710* tests of level II and IV had contextualized items with the listening texts, and the outcomes of levels. In other words, all the tests analyzed for this project are coherent with the goals of the levels.

4. Does the test contain items/tasks that are unambiguous to the test taker?

Before Specs

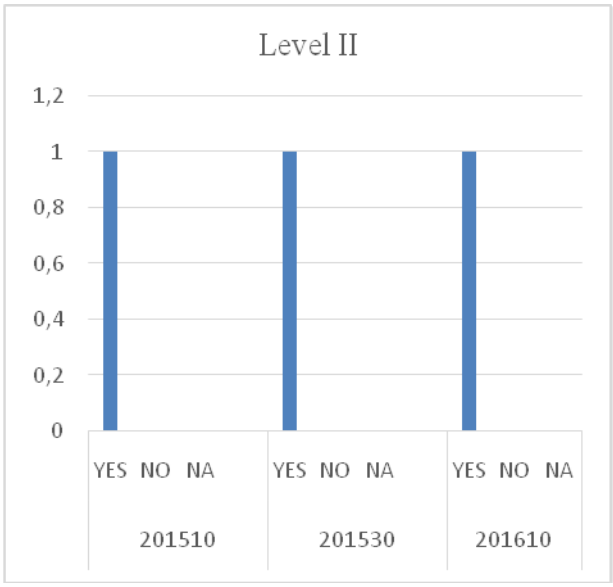


Figure 29

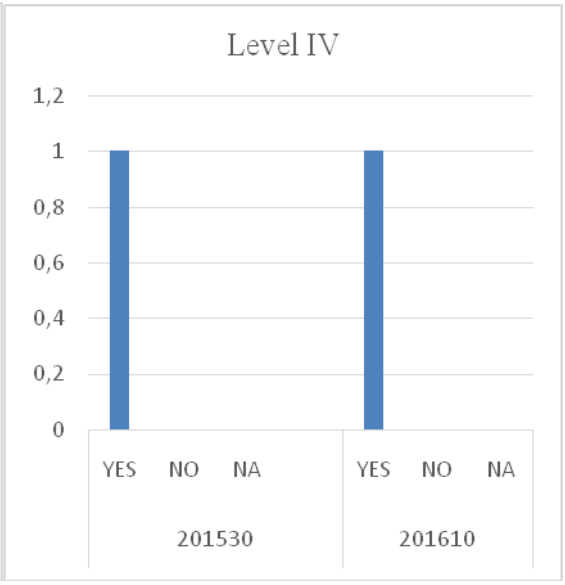


Figure 30

Figures 29 and 30 present whether the items and/or tasks in the tests are unambiguous to the test-taker or not. As illustrated in the figures, *201510*, *201530*, and *201610* tests of level II and *201530* and *201610* tests of level IV did not provide students with ambiguous items and/or tasks, understanding by unambiguous as not open to more than one interpretation (definition found in English Oxford Living Dictionary online). Furthermore, all the tests analyzed in this project propose to students clear and precise items and/or tasks.

See sample:

B. Listen for Details (QSKL1 CD2 Track 10; 2.44 mins)

Read the questions below, and then you will hear the presentation one more time. Circle the correct answer to complete each sentence. *(6 marks; 1 mark per question)*

1) The population of Cusco is about _____.
   a) 35,000
   b) 350,000
   c) 3,500,000

2) Machu Picchu is _____.
   a) a pretty city
   b) not near the mountains
   c) three hours from Cusco

3) The trip starts on ____.
   a) June 13th
   b) June 30th
   c) July 5th

4) The group is going to study Spanish for _____.
   a) two weeks
   b) three weeks
   c) four weeks

5) At the school, volunteers can ____.
   a) teach Spanish
   b) study music
   c) teach English

6) Volunteers say teaching children is _____.
   a) amazing
   b) enjoyable
   c) not fun

After Specs

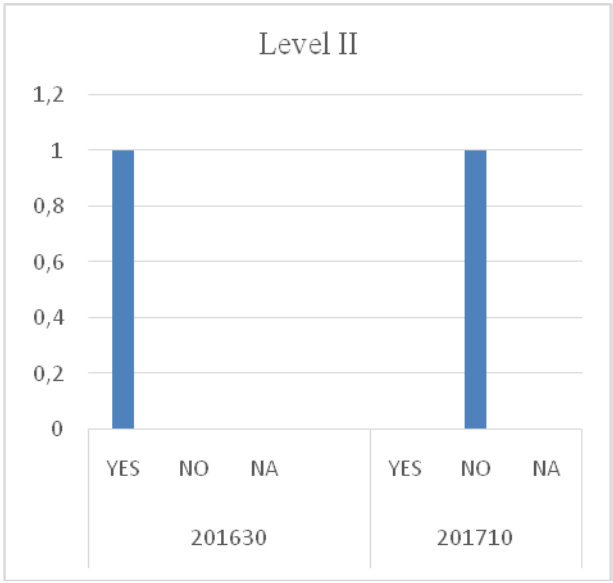


Figure 31

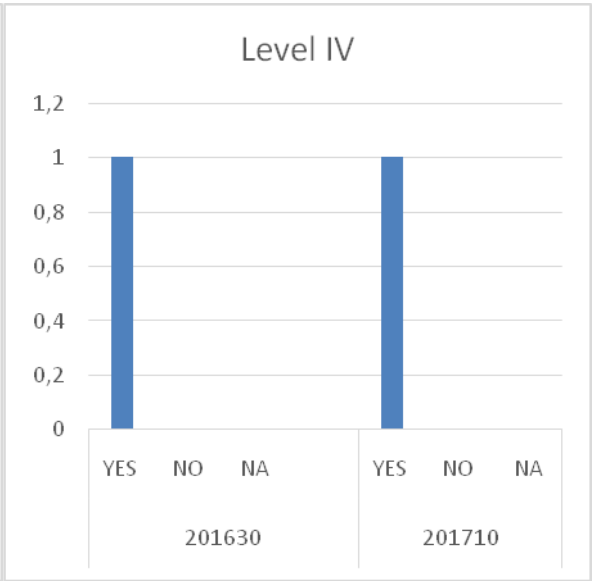


Figure 32

Figures 31 and 32 present whether the items and/or tasks in the tests are unambiguous to the test-taker or not. As illustrated in the figures, one of the two tests of level II *(201710)* provided students with ambiguous items and/or tasks. However, *201630* test of level II and *201630 and 201710* tests of level IV propose to students clear and precise items and/or tasks.

5. Is the test developed with a specific purpose and a particular group of test-takers?

Before Specs

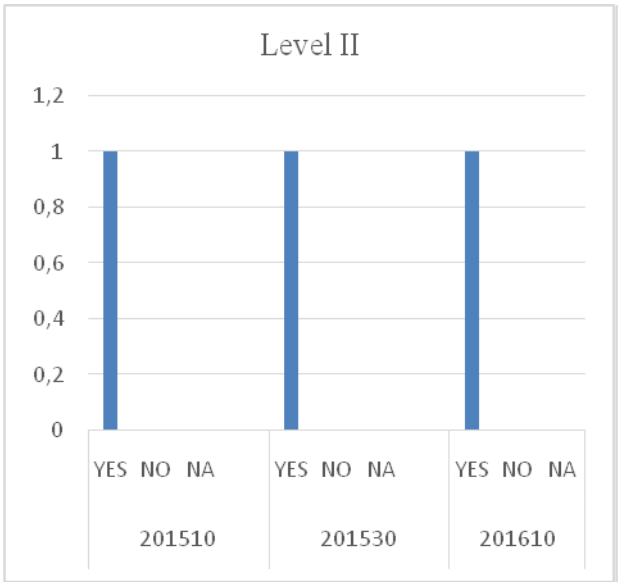


Figure 33

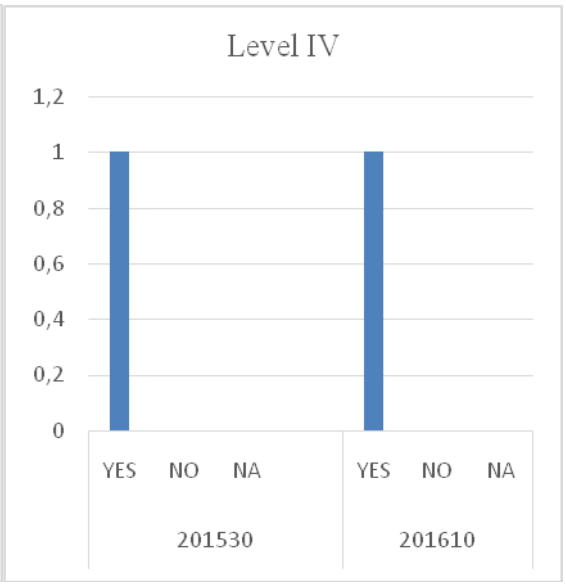


Figure 34

Figures 33 and 34 report information regarding the specific purpose and the particular groups of test-takers with which the tests are developed. It was observed that all the tests (*201510*, *201530*, and *201610* tests of level II and *201530* and *201610* tests of level IV) were designed bearing in mind a specific purpose and a group of test-takers. Likewise, the outcomes of each level, the conditions of the students taking the tests and the topics of the books are reflected in the listening texts chosen and the items included. (See samples)

Level II

Part One

A. Listen for Main Ideas (QSKL1 CD2 Track 10; 2.44 mins)

Volunteer Vacations is a travel company that offers work and travel around the world. You are going to listen to the owner of this company give a presentation about jobs for volunteers in Cusco, Peru. Read the items and then listen to the information. You will hear it one time. Check (✓) the six things that the volunteers are going to do. *(6 marks; 1 mark per question)*

1) work on a farm ____
2) visit Machu Picchu ____
3) study Spanish ____
4) visit museums ____
5) live with a host family ____
6) relax on the beach ____
7) help sick people ____
8) teach at a school ____
9) repair a school ____
10) learn about Peruvian culture ____

As the examples show, the exams have a specific purpose which is to evaluate the outcomes of the level and a specific group of test-takers, students of levels II and IV.

After Specs

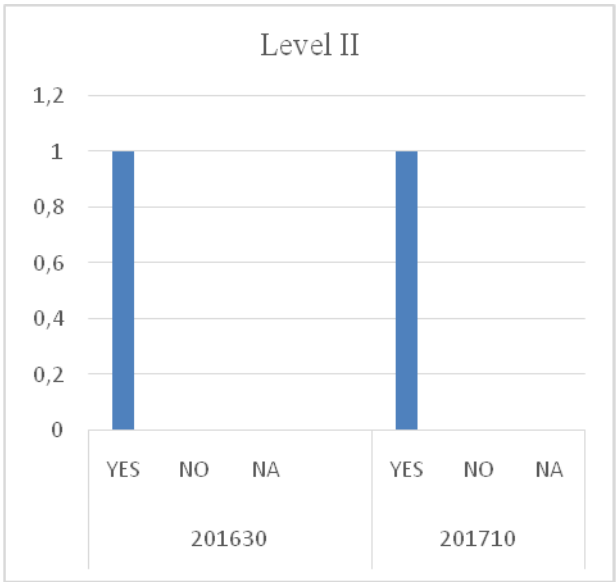


Figure 35

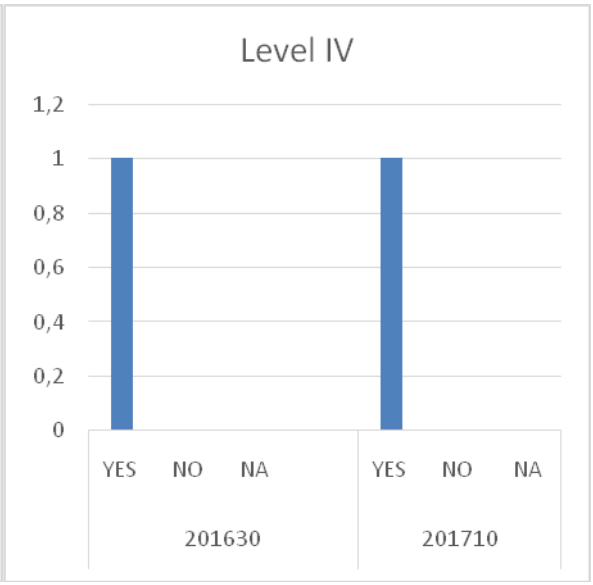


Figure 36

Figure 35 and figure 36 report information regarding the specific purpose, the particular groups of test-takers, and the specific language use with which the tests are developed. It is quite clear that all *201630 and 201710* tests of level II and all *201630 and 201710* tests of level IV were designed bearing in mind a specific purpose, a group of test-takers, and a specific language use. In this way, the outcomes of each level, the conditions of the students taking the tests and the topics of the books are reflected in the listening texts chosen and the items included (See samples below).

Level II

Universidad del Norte – Instituto de Idiomas
Listening Assessment
Northstar 2B (L&S Units 5 & 7)

Exam Code: 201630A

INSTRUCTIONS
Write your name and Exam Code on the Answer Sheet.
Read the instructions for each part carefully.
Answer all questions.

Write your answers on the answer sheet.

Part One: Listening One: "Business is a Game"

A. Listen for Main Ideas (QSKL2 CD1 Track 40; 4.28 mins)
Two friends, Moy and Hannah, are talking about an assignment for a business class. The assignment is to play a computer game that teaches some business ideas. You will listen to their conversation one time. Circle the correct answer for each question **on the Answer sheet**. *(4 points; 1 point per question)*

1) What does Moy think about the Lemonade Game?

a) It´s fun, but it can´t help him learn about business.
b) It isn´t very interesting, but it can teach him about business.
c) It´s entertaining and useful for learning about business.

2) Which of these things can you learn from the Lemonade Game?

a) the connection between supply and demand
b) the connection between supply and price
c) the connection between lemons and supply.

Level IV

ELP IV LISTENING EXAM ---TEST BOOKLET---

Universidad del Norte – Instituto de Idiomas
LISTENING ASSESSMENT 201710
NorthStar 3B (L&S Units 6 and 7)
SECTION ONE: LISTENING COMPREHENSION
_____/40 POINTS

LISTENING ONE: THE GLOBAL CHANGE EFFECT [20 points]

Main Ideas
a. Read the questions and choose the correct answer below. (2 points each, 4 points total)

1. According to the conversation, we can say that…
a. There is a lot of confusion about climate change.
b. People know exactly what to expect from climate change.
c. Science is not helping much when it comes to informing about climate change

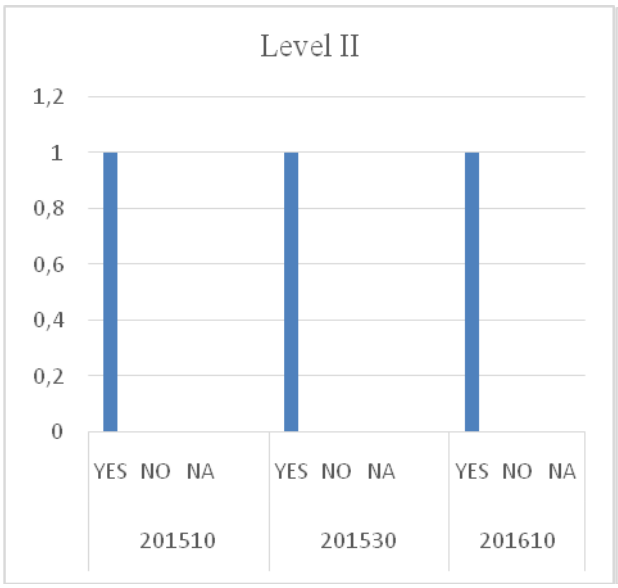
2. According to Matt, we can say that he…
a. Is not convinced that humans are the main cause of climate change.
b. Wants to convince Kate that climate change is dangerous.
c. Believes that climate change is an effect of evolution.

In the examples, it can be seen that the exams have a specific purpose, evaluate the level outcomes and they also have a particular group of test-takers, students of levels II and IV.

Test Items Quality

1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)

Before Specs



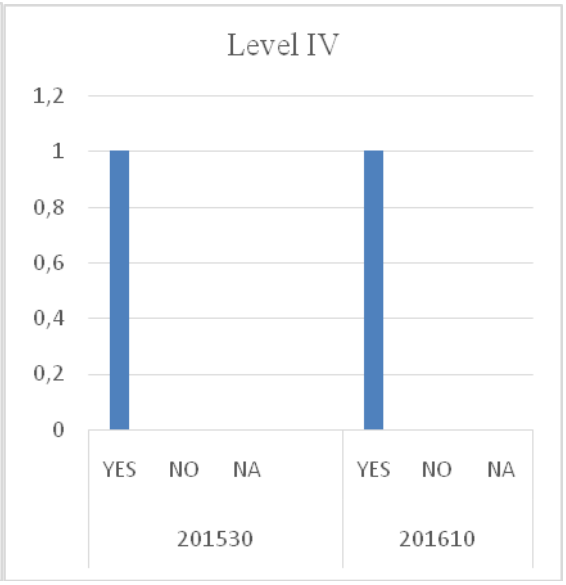


Figure 37 Figure 38

Figures 37 and 38 show the clarity of the stem whether or not it clearly indicates what students have to do in the test. As it can be seen, *201510, 201530,* and *201610* tests of level II accurately direct what students must do. Similarly, *201530* and *201610* tests of level IV also provide students with suitable instructions that allow them do the exercises. In this way, all tests seem to be clear and precise at the moment of giving directions.

After Specs

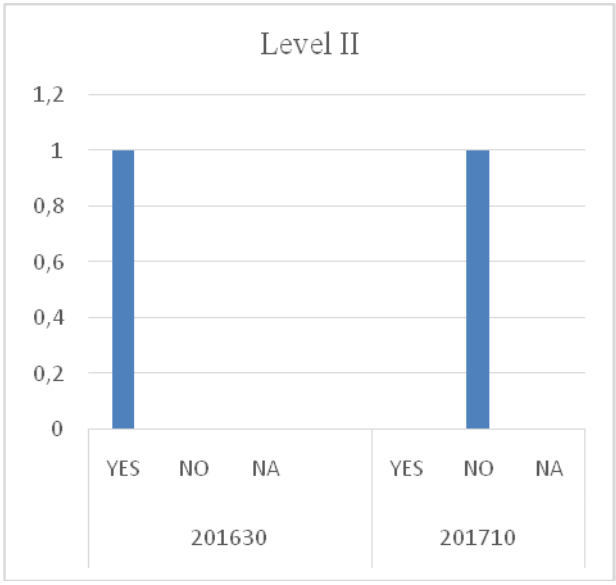


Figure 39

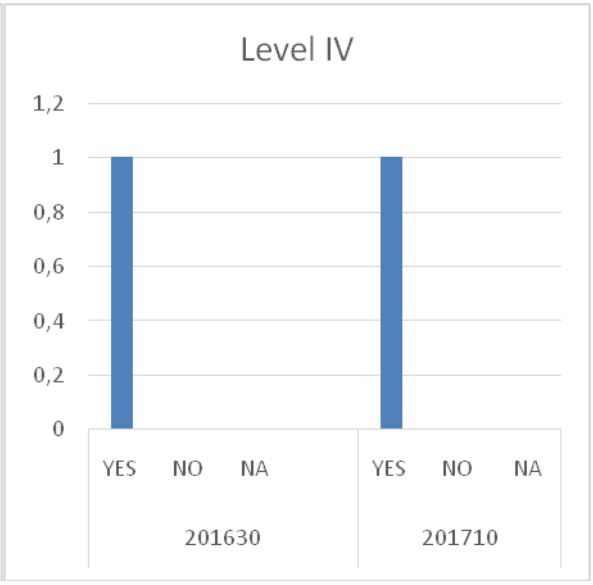


Figure 40

Figures 37 and 38 show the clarity of the stem whether or not it clearly indicates what students have to do in the test. As it can be seen, one out of the two tests of level II *(201710)* did not direct what students must do. However, the other level II listening tests and *201630 and 201710* tests of level IV provide students with suitable instructions that allow them do the exercises. In this way, most tests seem to be clear and precise at the moment of giving directions.

2. Is each option clearly identified as the answer to the question asked?

Before Specs

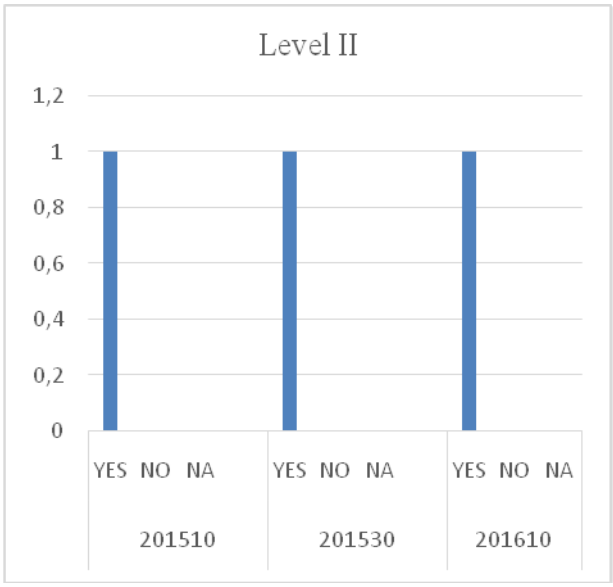


Figure 41

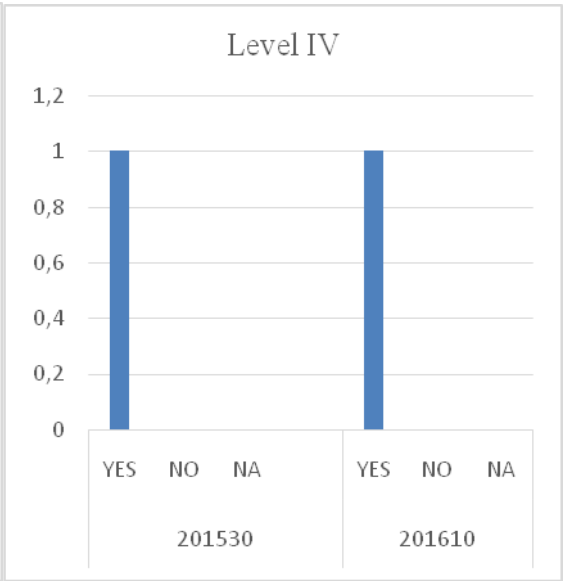


Figure 42

Figure 41 and 42 illustrate if, in multiple choice questions, each option is clearly identified as the answer to the question asked. *201510, 201530* and *201610* test of level II present clear options that can be distinguished as the answers of the questions of the tests. Likewise, *201530* and *201610* of level IV also show that these options are easily recognized as the possible answers. Therefore, each test of both levels appears to have identified the options to the questions formulated in the exams.

Part One (Q: Skills for Success. Listening and Speaking 3 (Class Audio Set)

Listen for Main Ideas (4 points)

Quickly read the questions. Now listen to a class discussion between a teacher and some teenage students between the ages of 16 and 18. You will hear it one time. Check (✓) the **BEST** answer.

1 What would make a good title for this class discussion? ○ Individual Responsibility ○ Too Young to Be Responsible ○ Parental Responsibility	3 What are parental expectations according to these young people? ○ parents want to know where they are ○ parents don't want their children to lie ○ parents give them too many responsibilities
2 What bothers most of these young people? ○ their parents are not always at home ○ their parents don't trust them ○ their parents bother them all the time	4 We learn that _____. ○ everyone has the same responsibilities at home ○ young people lie to their parents and are not responsible ○ not everyone agrees about being independent

After Specs

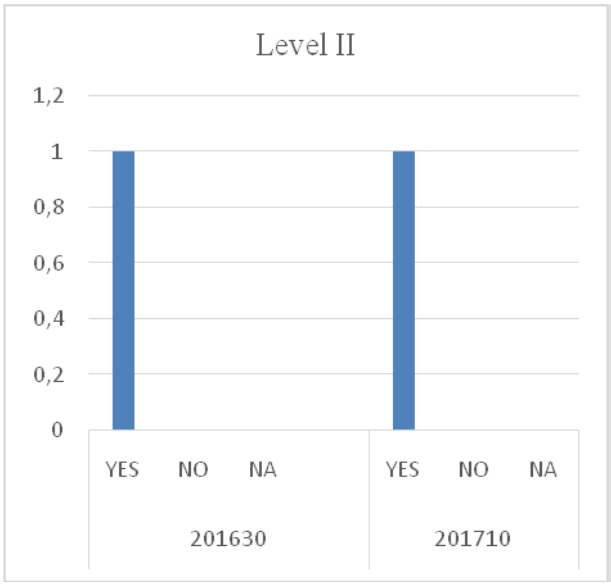


Figure 43

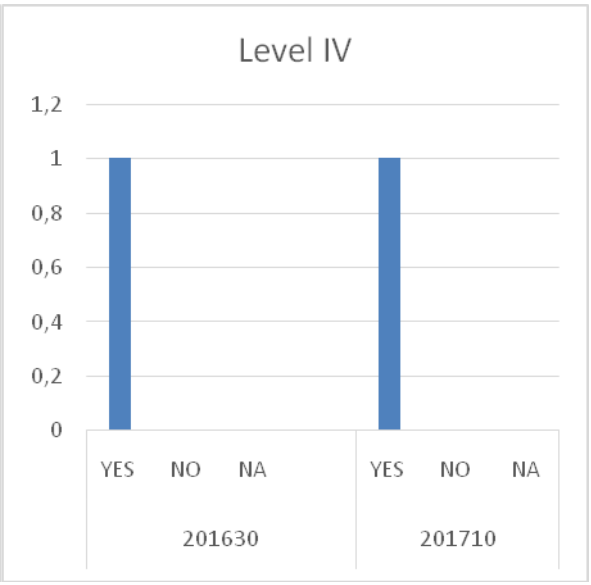


Figure 44

Figure 41 and 42 illustrate if, in multiple choice questions, each option is clearly identified as the answer to the question asked. *201630 and 201710* tests of level II and *2016030 and 201710*

tests of level IV presented clear options that can be distinguished as the answers of the questions of the tests. Likewise, each test of both levels appears to have identified the options to the questions formulated in the exams.

3. Is the answer to each question text dependent (Audio)? (it does not depend on students' prior knowledge)

Before Specs

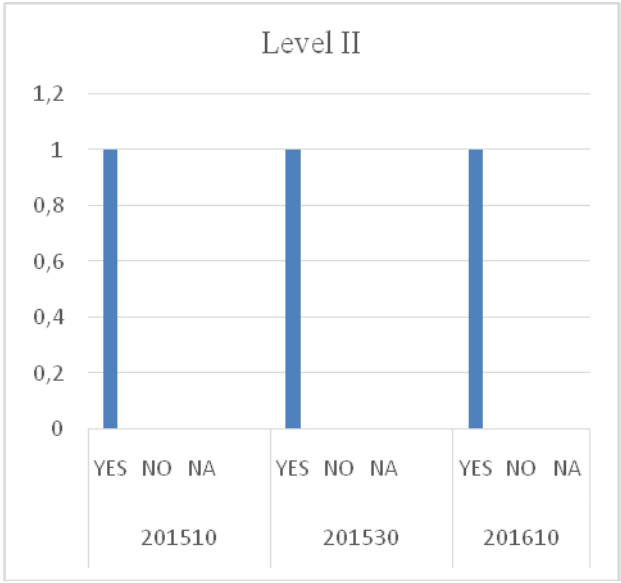


Figure 45

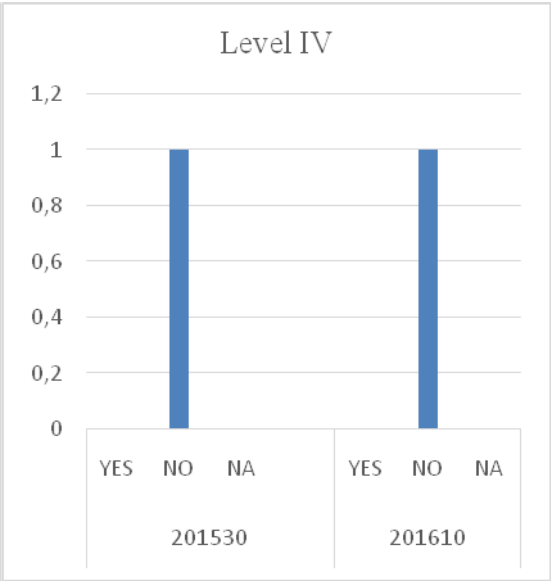


Figure 46

Figure 45 and figure 46 read the listening text dependency that answers of questions have. *201510, 201530* and *201610* tests of level II revealed that all answers are text dependent. This means that students really need to listen to the audios if they want to choose the correct answer. In contrast, the answers in *201530* and *201610* tests of level IV are not text dependent because there is no need to listen to the audios to spot the right answer since other options are

quite obvious. Hence, while tests of level II possess dependent answers to the text, tests of level IV do not do so since students can actually answer questions by their prior knowledge and not by listening to the text itself.

Level II Sample of text dependent questions.

B. Listen for Details (QSKL2 CD1 Track 37; 3.23 mins)

Read the questions, and then you will hear the speech one more time. Circle the correct answer. *(5 points; 1 point per question)*

1) When did Bradley start the Milton Bradley Company?

a) 1960 b) 1911 c) 1860

2) What colors were the squares on the board for *The Checkered Game of Life*?

a) red and blue b) red and black c) blue and black

3) How many copies did *The Checkered Game of Life* sell in the first year?

a) over 4,000 b) over 40,000 c) over 400,000

Sample of not text dependent questions

Listen for Main Ideas (2 points)

Listen carefully to this brief announcement about prenuptial agreements. Check (✓) the **BEST** answer that completes each statement.

1 The purpose of a prenuptial agreement is to _____ a family. ○ assess ○ protect ○ worry	2 A prenuptial agreement is a _____ between spouses. ○ budget ○ contract ○ expectation

Level IV

Listen for Main Ideas (2 points)

Listen carefully to this brief announcement about prenuptial agreements. Check (✓) the **BEST** answer that completes each statement.

1 The purpose of a prenuptial agreement is to _____ a family. ○ assess ○ protect ○ worry	2 A prenuptial agreement is a _____ between spouses. ○ budget ○ contract ○ expectation

After Specs

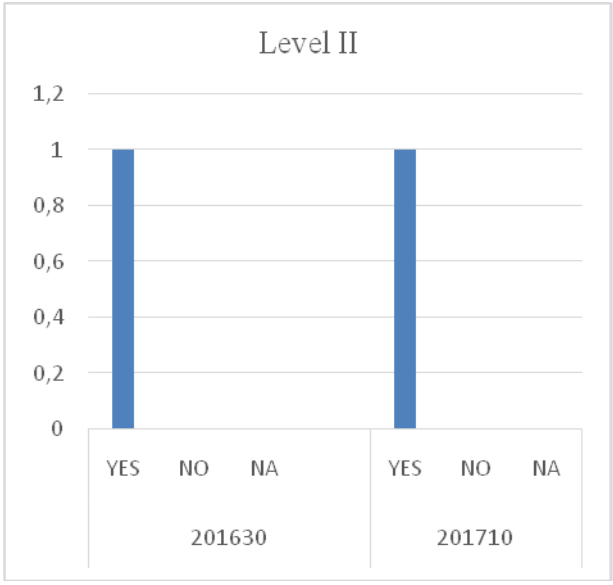


Figure 47

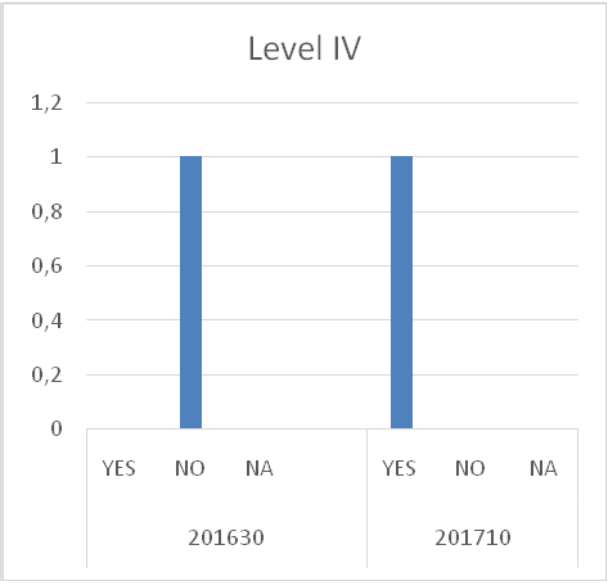


Figure 48

Figure 47 and figure 48 read the listening text dependency that answers to questions have.

201630 and *201710* tests of level II and *201710* test of level IV revealed that all answers are text dependent. This means that students really need to listen to the audios if they want to choose the correct answer. In contrast, the answers in *201630* test of level IV are not text dependent because

there is no need to listen to the audios to spot the right answer since other options are quite obvious.

Level II

Part Two: Listening Two: Doctor's opinion

A. Listen for Main Ideas (NS2AT CD1 Track 40)

Read the questions. Listen to a conversation between Mary Ann and her doctor. You will hear it one time. Circle the correct answer for each question **on the Answer sheet**. (4 points)

1. Why does Mary Ann visit her doctor?
   a) She has a physical exam every year.
   b) She feels bad.
   c) She looks thin.

2. How much exercise should Mary Ann get?
   a) Thirty minutes or more a day.
   b) Thirty minutes every other day.
   c) About an hour twice a week.

3. What is the doctor's opinion of Thin Fast?
   a) She thinks it's a healthy way to lose weight.
   b) She doesn't think it has side effects.
   c) She doesn't think it's safe.

4. The Doctor hopes that in a month Mary Ann…
   a) Will weigh less.
   b) Will be more energetic.
   c) Won't be exercising so much.

Level IV

Details
Circle who says each statement. (2 points per question; 8 points total)

1. We don't have enough statistics to be sure.

a) Kate
b) Matt
c) Nobody .

2. You know they've said sea levels are going to rise by quite a few meters over the next fifty to a hundred years.

a) Kate
b) Matt
c) Nobody

3. The loss of the Amazon rainforest could cause problems.

a) Kate
b) Matt
c) Nobody

4. People only go to zoos to see the very rare, endangered animals that are there.

a) Kate
b) Matt
c) Nobody

4. Is the answer to each question text dependent (it can only be answered by the information provided by the text – it does not depend on other stems or keys)?

Before Specs

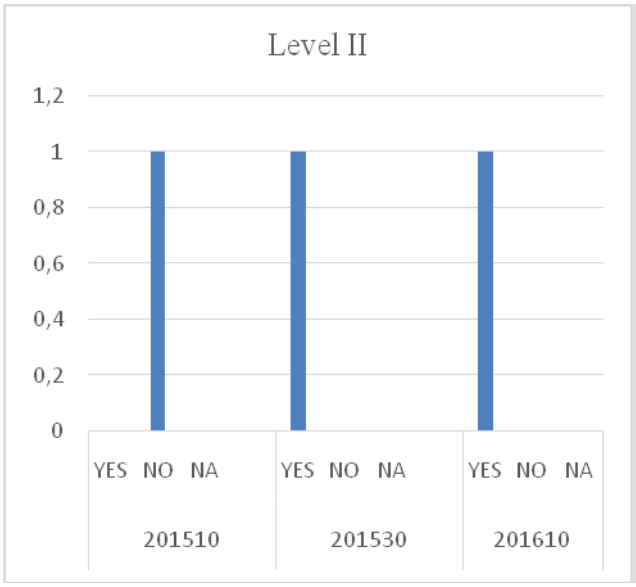


Figure 49

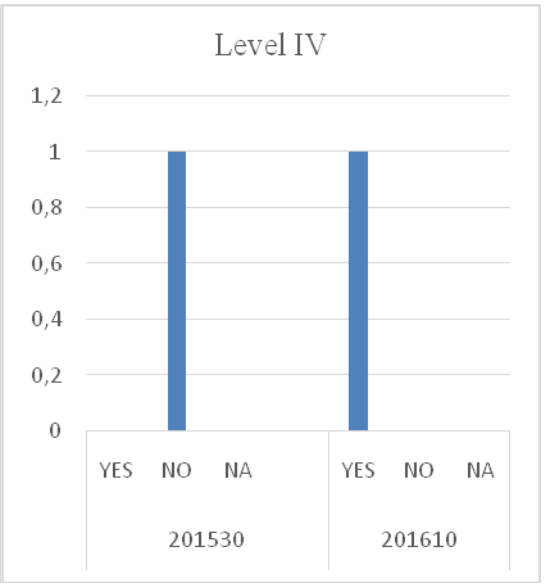


Figure 50

Figure 49 and figure 50 shows the text dependency that answers to questions have. Answers in *201510* tests of level II are not text dependent because they can be answered with the information of other stems and keys rather than the listening text itself. On the contrary, *201530* and *201610* tests of level II have text dependent answers due to the fact that listening to the audio is compulsory if the right question is intended to be answered. Additionally, while *201530* test of level IV does not have text dependent answers, *201610* does. As shown in the figures, text dependency answers of both levels, II and IV, still have certain drawbacks that affect the quality of questions in a test.

Level II

B. Listen for Details (QSKL2 CD1 Track 37; 3.23 mins)

Read the questions, and then you will hear the speech one more time. Circle the correct answer. *(5 points; 1 point per question)*

1) When did Bradley start the Milton Bradley Company?

a) 1960 b) 1911 c) 1860

2) What colors were the squares on the board for *The Checkered Game of Life*?

a) red and blue b) red and black c) blue and black

3) How many copies did *The Checkered Game of Life* sell in the first year?

a) over 4,000 b) over 40,000 c) over 400,000

As observed in the sample, students really need to listen to the audios if they want to spot the correct answer. Otherwise, they will not be able to identify the correct answer by their prior knowledge.

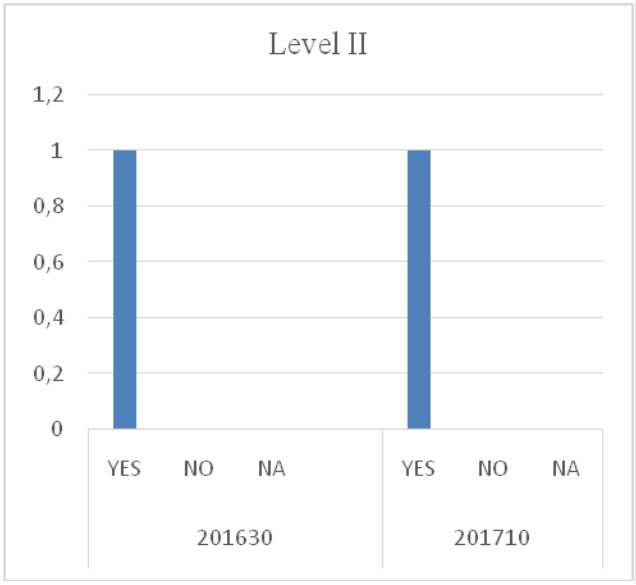
Level IV

Listen for Main Ideas (2 points)

Listen carefully to this brief announcement about prenuptial agreements. Check (✓) the **BEST** answer that completes each statement.

1 The purpose of a prenuptial agreement is to _____ a family. ○ assess ○ protect ○ worry	2 A prenuptial agreement is a _____ between spouses. ○ budget ○ contract ○ expectation

After Specs



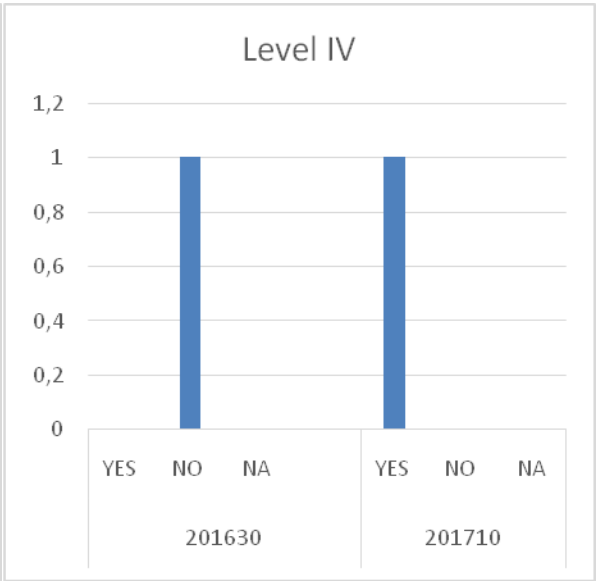


Figure 51 Figure 52

Figure 51 and figure 52 indicate the text dependency that answers to questions have. Answers in *201630* test of level IV are not text dependent because they can be answered with the information of other stems and keys rather than the listening text itself. On the contrary, *201630 and 201710* tests level II and *2011710* test of level IV have text-dependent answers due to the fact that listening to the audio is compulsory if the right question is intended to be answered. In other words, most of the listening tests after Test Specs comply with the quality of test items.

Level II

Part Two: Listening Two: Doctor's opinion

A. Listen for Main Ideas (NS2AT CD1 Track 40)

Read the questions. Listen to a conversation between Mary Ann and her doctor. You will hear it one time. Circle the correct answer for each question **on the Answer sheet**. (4 points)

1. Why does Mary Ann visit her doctor?
   a) She has a physical exam every year.
   b) She feels bad.
   c) She looks thin.

2. How much exercise should Mary Ann get?
   a) Thirty minutes or more a day.
   b) Thirty minutes every other day.
   c) About an hour twice a week.

3. What is the doctor's opinion of Thin Fast?
   a) She thinks it's a healthy way to lose weight.
   b) She doesn't think it has side effects.
   c) She doesn't think it's safe.

4. The Doctor hopes that in a month Mary Ann...
   a) Will weigh less.
   b) Will be more energetic.
   c) Won't be exercising so much.

As observed in the sample, students really need to listen to the audios if they want to spot the correct answer. Otherwise, they will not be able to identify the correct answer by their prior knowledge.

Level IV

Details
Circle who says each statement. (2 points per question; 8 points total)

1. We don't have enough statistics to be sure.

a) Kate
b) Matt
c) Nobody .

2. You know they've said sea levels are going to rise by quite a few meters over the next fifty to a hundred years.

a) Kate
b) Matt
c) Nobody

3. The loss of the Amazon rainforest could cause problems.

a) Kate
b) Matt
c) Nobody

4. People only go to zoos to see the very rare, endangered animals that are there.

a) Kate
b) Matt
c) Nobody

5. Are the options of the questions parallel? (the same formatting of speech parts)

Before Specs

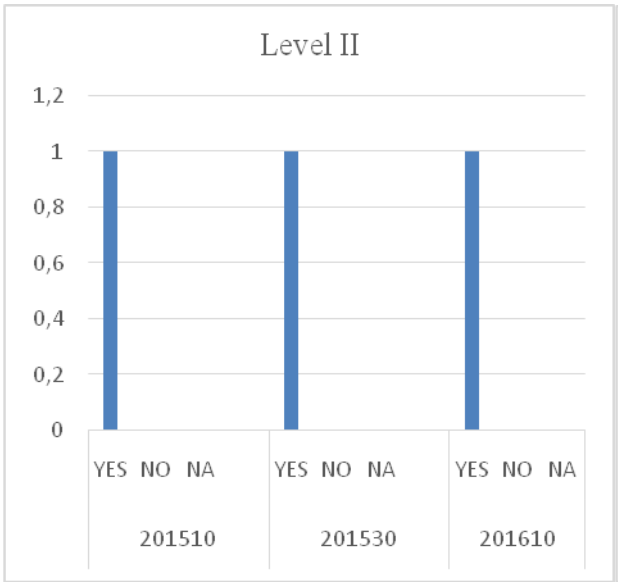


Figure 53

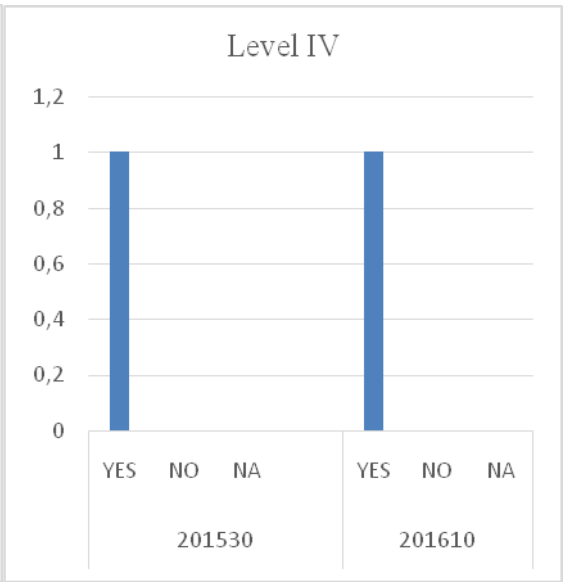


Figure 54

Figure 53 and figure 54 suggest the parallelism of the options of the questions. *201510*, *201530* and *201610* tests of level II present parallel options to the questions since they start with the same parts of the speech (nouns, verbs, adjectives). Correspondingly, *201530* and *201610* also present parallel options to the questions of tests due to the fact they follow the same formatting or speech parts. Thus, all tests in both level II and IV have parallel choices of questions asked in the tests.

After Specs

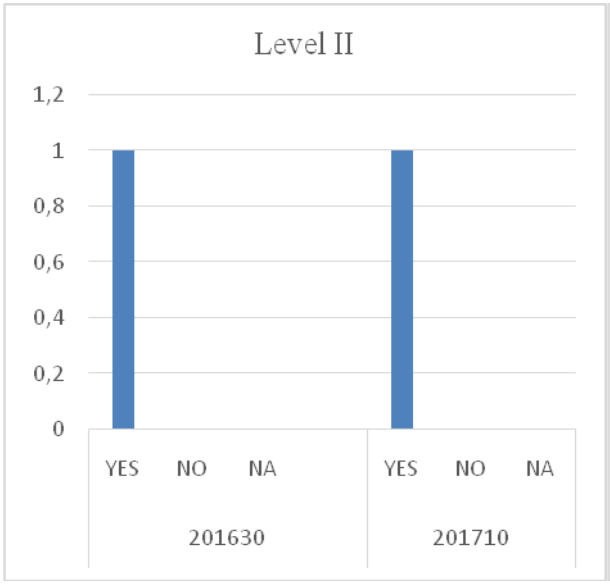


Figure 55

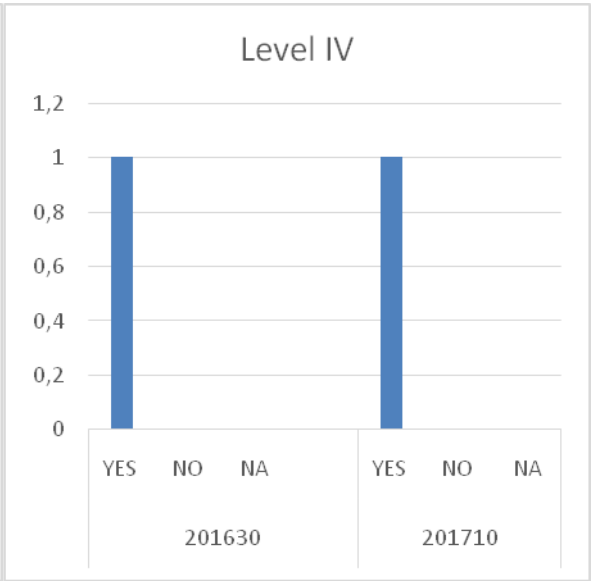


Figure 56

Figure 53 and figure 54 indicate the parallelism of the options of the questions. All *201630* and *201710* test of level II and all *201630 and 201710* tests of level IV present parallel options to the questions since they start with the same parts of the speech (nouns, verbs, adjectives). To be more precise, all listening tests that were designed after Test Specs present parallel options to the questions of tests due to the fact they follow the same formatting or speech parts.

6. Are the questions formulated as affirmative statements?

Before Specs

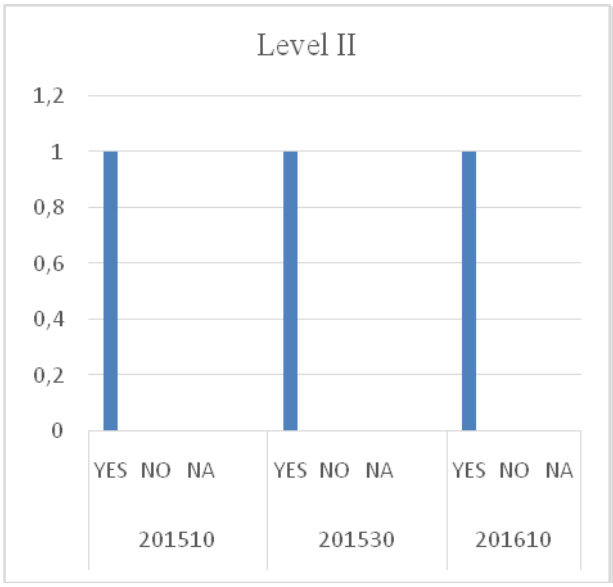


Figure 57

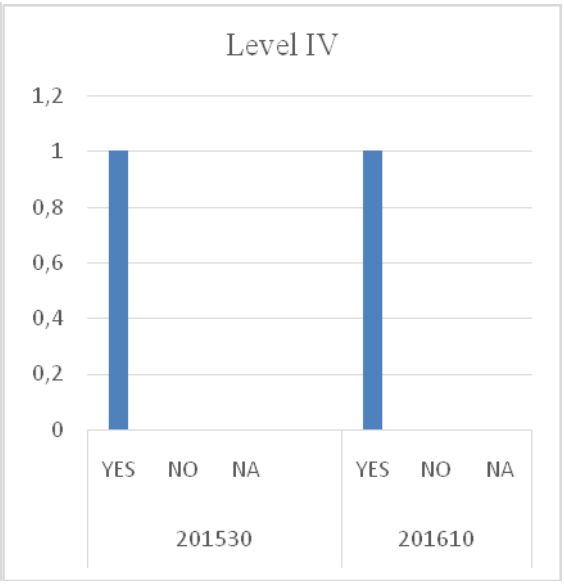


Figure 58

Figure 57 and figure 58 indicate whether the questions were formulated affirmatively or not. This means that having these kinds of questions for students to answer in a test presents a positive impact due to the fact that guidelines instructions were followed. 201510, 201530 and 201610 had affirmative questions as well as those 201530 and 201610 tests which also possessed affirmative ones. Consequently, in 201530 and 201610, all the items in the test were also formulated as affirmative statements. Furthermore, all the tests in both level II and IV asked questions affirmatively. Formulating affirmative statements is important to avoid confusing students because negative statements might lead to students' confusion if these have not been properly practiced in class.

After Specs

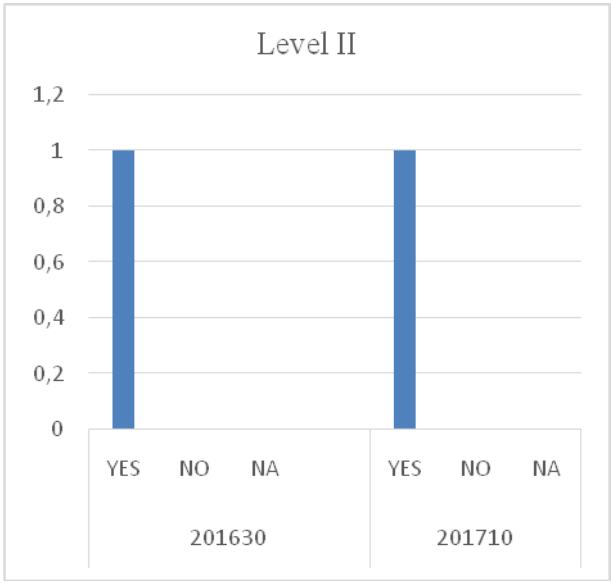


Figure 59

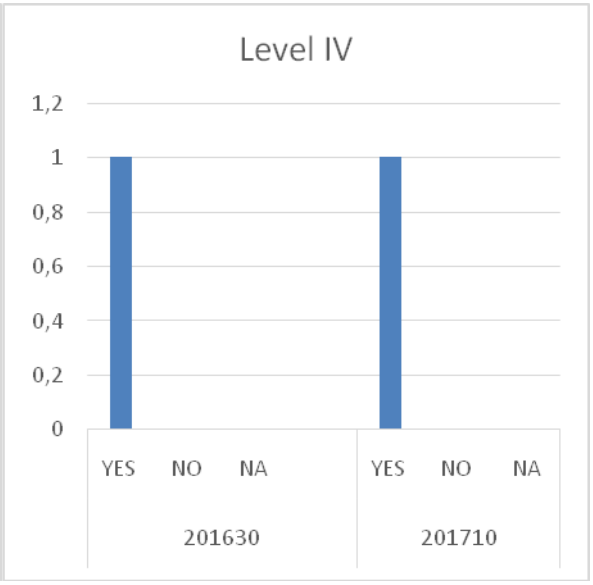
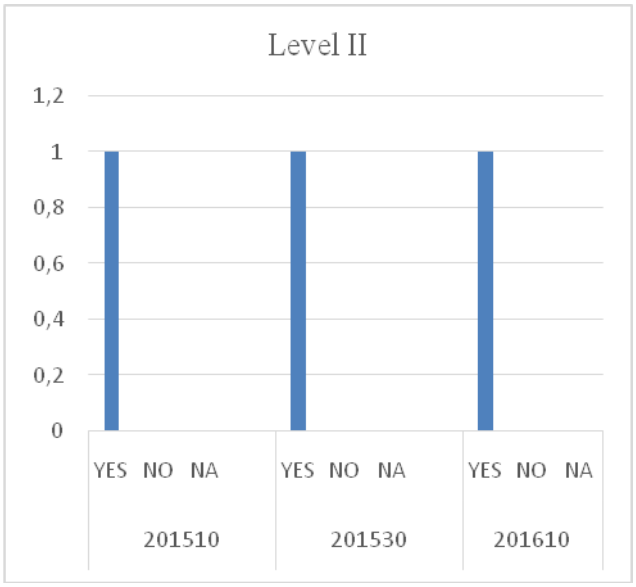


Figure 60

Figure 59 and figure 60 indicate whether the questions were formulated positively or not. This means that having these kinds of questions for students to answer in a test presents a positive impact due to the fact that Test Specs instructions were followed. All *201630 and 201710* tests of level II had affirmative questions. Correspondingly, all the items in *201630 and 201710 tests* of level IV were also formulated positively. Hence, all the tests in both level II and IV asked positive questions. Formulating affirmative statements is important to avoid confusing students because negative statements might lead to students' confusion if these have not been properly practiced in class.

7. Are distractors well designed?

Before Specs



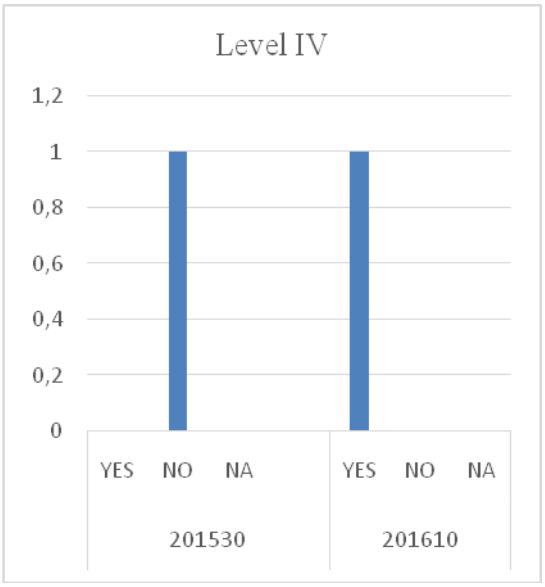


Figure 61 Figure 62

Figures 61 and 62 present the design of distractors in multiple choice questions. Distractors are very important in multiple choice questions because they do not only challenge students to identify what the correct answer is, but they also foster critical analysis in the application of the exams. As presented in the figures, *201510*, *201530*, and *201610* tests of level II and *201610* test of level IV distractors are designed well enough to comply with their objective. On the contrary, in *201530* test of level IV the design of the distractors was not good enough. This means that in all the tests of level II examined for this study and the *201610* test of level IV, the keys that were not the correct answers were not too obvious for students to immediately guess they were false. However, in the *201530* test of level IV, the keys that were incorrect were also too obvious for students. They could easily guess the right answer not because they knew it was correct, but because the incorrect ones were evident. (See questions 1 and 2 in the sample below)

Sample

B. Listen for Details (QSKL1 CD2 Track 10; 2.44 mins)

Read the questions below, and then you will hear the presentation one more time. Circle the correct answer to complete each sentence. *(6 marks; 1 mark per question)*

1) The population of Cusco is about _____.
   a) 35,000
   b) 350,000
   c) 3,500,000

2) Machu Picchu is _____.
   a) a pretty city
   b) not near the mountains
   c) three hours from Cusco

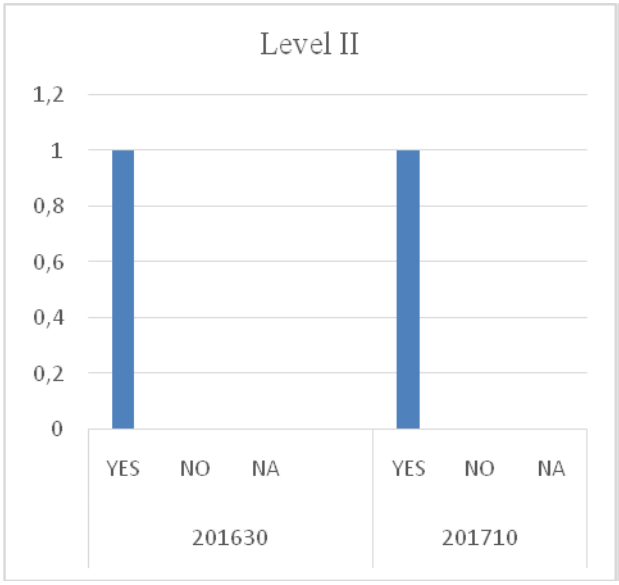
3) The trip starts on ____.
   a) June 13th
   b) June 30th
   c) July 5th

4) The group is going to study Spanish for _____.
   a) two weeks
   b) three weeks
   c) four weeks

5) At the school, volunteers can ____.
   a) teach Spanish
   b) study music
   c) teach English

6) Volunteers say teaching children is _____.
   a) amazing
   b) enjoyable
   c) not fun

After Specs



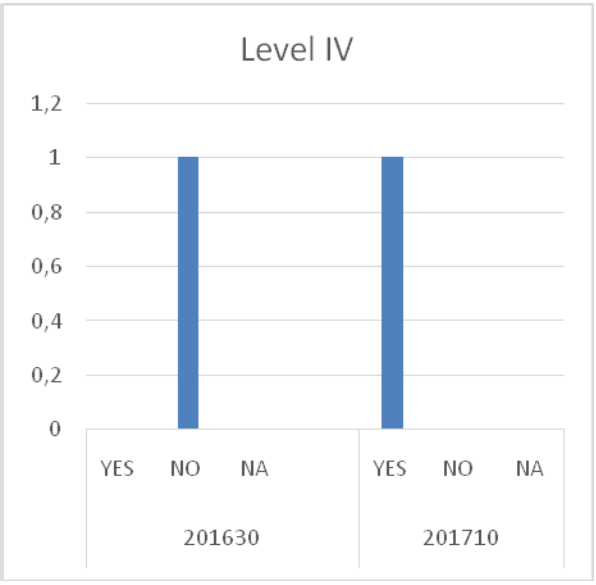


Figure 63 Figure 64

Figures 63 and 64 show the design of distractors in multiple choice questions. Distractors are very important in multiple choice questions because they do not only challenge students to

identify what the correct answer is, but they also foster critical analysis in the application of the exams. *201630* listening test did not have well designed distractors that challenged students. However, *201630* and *201710* tests of level II and *201710* test of level IV contained well designed distractors that complied with the quality of test items. (See sample)

Sample

Part One: Listening One: "Business is a Game"

A. Listen for Main Ideas (QSKL2 CD1 Track 40; 4.28 mins)

Two friends, Moy and Hannah, are talking about an assignment for a business class. The assignment is to play a computer game that teaches some business ideas. You will listen to their conversation one time. Circle the correct answer for each question **on the Answer sheet**. *(4 points; 1 point per question)*

1) What does Moy think about the Lemonade Game?

a) It´s fun, but it can´t help him learn about business.
b) It isn´t very interesting, but it can teach him about business.
c) It´s entertaining and useful for learning about business.

2) Which of these things can you learn from the Lemonade Game?

a) the connection between supply and demand
b) the connection between supply and price
c) the connection between lemons and supply.

3) What happened when Hannah played the game?

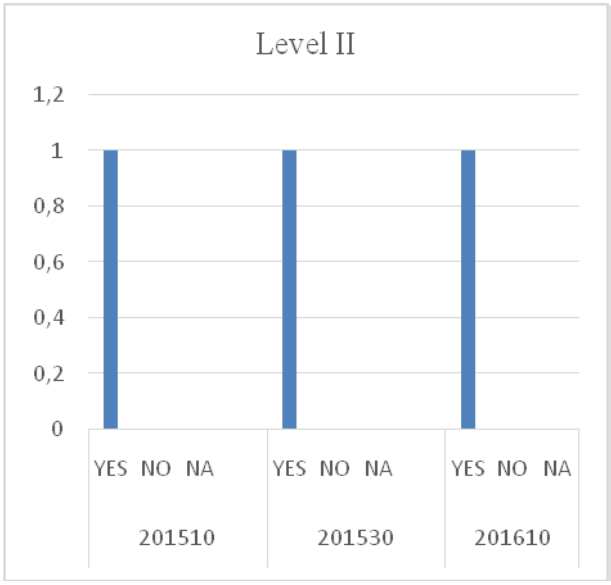
a) She made a lot of money.
b) She lost a little money.
c) She made too much lemonade.

4) What is Hannah´s opinion of using a game to learn business?

a) She thinks it is a good way to learn.
b) She thinks it only works for lemonade businesses.
c) She thinks it is not the best idea for a university class.

1. Are the numbers of questions in chronological order?

Before Specs



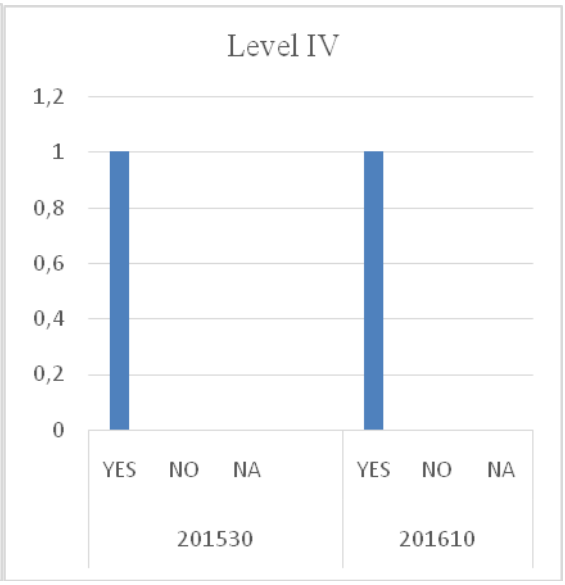


Figure 65 Figure 66

Figures 65 and 66 indicate if the number of questions is in a correct numerical order. As presented in the figure, the items of the *201510*, *201530*, and *201610* tests of level II and *201530* and *201610* tests of level IV are constructed following a specific chronological order. In this way, all the tests studied for this project are conformed by questions that are formulated in a coherent chronological order. Numerical order is important to avoid confusing students because it gives consistency and makes tests easier to follow.

After Specs

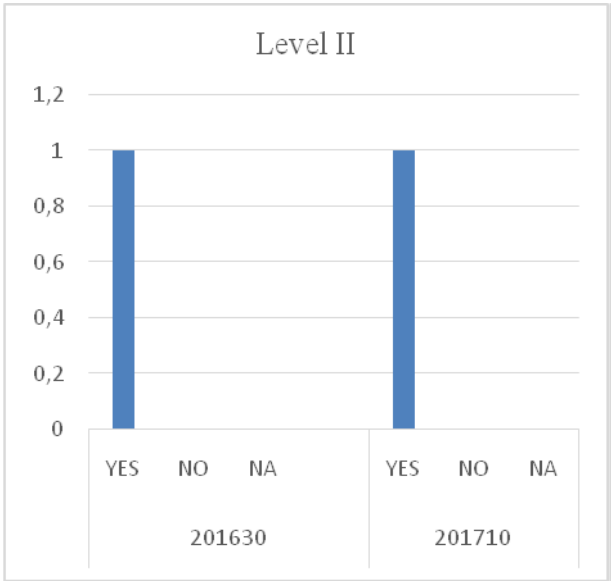


Figure 67

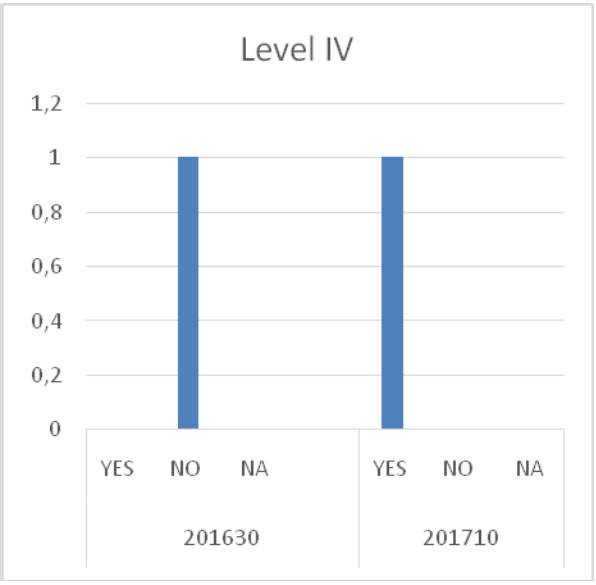


Figure 68

Figures 67 and 68 show whether tests, after test specs, follow a specific chronological order. *201630* and *201710* tests of level II and *201710* test of level IV had items that were constructed following a specific numerical order while *201630* test of level IV did not have questions that are formulated in a coherent numerical order. Numerical order is important to avoid confusing students because it gives consistency and makes tests easier to follow.

9. Do matching exercises have at least two extra options?

Before Specs

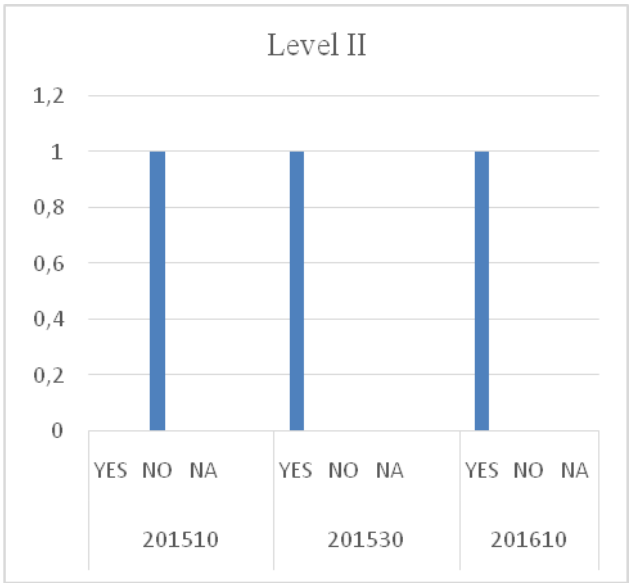


Figure 69

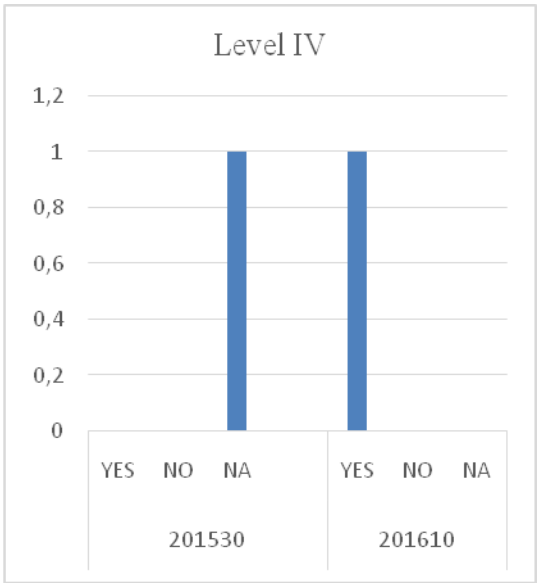


Figure 70

Figures 69 and 70 present whether matching exercises had two extra options or not. Having extra options in matching exercises is important to prevent students answering items by process of elimination or it helps to avoid double penalization when they make a mistake. As it can be observed in the figures, regarding level II, *201510* test did not have extra options in matching exercises, while *201530*, and *201610* tests and *201610* test of level IV did have two extra choices. On the other hand, the *201530* test of level IV did not have any matching exercises. That is to say, in three out of the five tests analyzed for this study, the matching exercises had extra options. (See Sample)

Sample:

VOCABULARY: **_____/ 10 POINTS**

A. **Match the boldfaced words and phrases on the left with their definitions on the right. Write the letter of the definition next to each word or phrase. Not all of the definitions will be used.**

___ 1. bother	**A. able to change easily**
___ 2. budget	**B. belief that something will happen**
___ 3. check up on	**C. examine if someone is doing things**
___ 4. expectation	**D. a legal agreement**
___ 5. quirk	**E. a plan for how to spend money**
	F. annoy
	G. unusual trait

After Specs

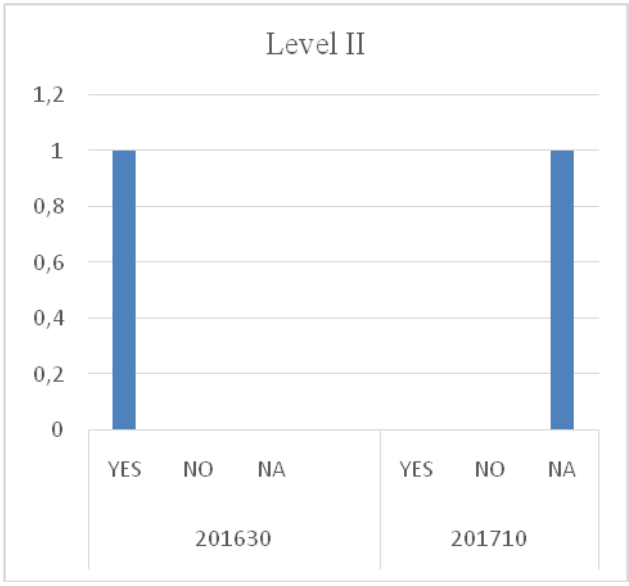


Figure 71

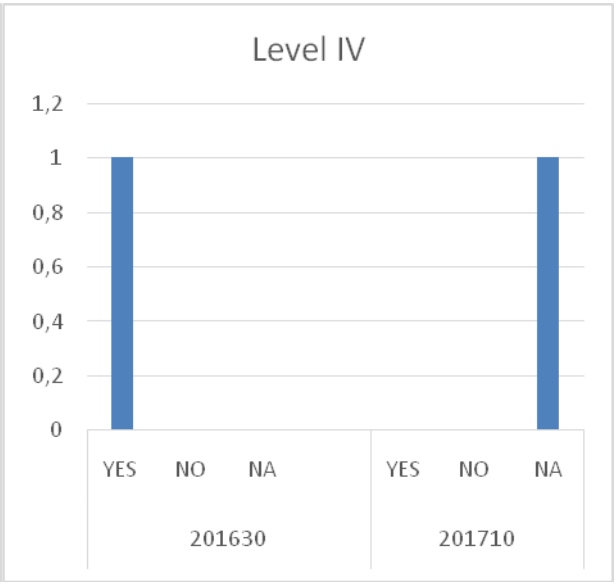


Figure 72

Figures 71 and 72 show whether matching exercises had two extra options or not. Having extra options in matching exercises is important to prevent students answering items by process of elimination. *201630* listening tests of level II and level IV had matching exercises with two extra options but *201710* tests of level II and level IV did not even have matching exercises. This

means that 50% of listening tests in both levels possessed matching exercises with two extra options. (See sample)

Sample

Part Three: Vocabulary

A. Match the words to their definitions. Write the letter **on the Answer sheet**. There are two extra options. *(10 points; 1 point per question)*

A	discourage	**1)** to remove or throw away something you do not want
B	couch potato	**2)** money people are required to pay the government
C	treatment	**3)** to give someone a reason for doing something
D	claim	**4**) sickness or disease
E	criticize	**5)** someone who spends a lot of time sitting and watching TV
F	needle	**6)** to say that something is true, even though it may not be
G	deal with	**7)** to suggest that someone not do something
H	consumption	**8)** something that is done to help someone who is ill or injured
I	tax	**9)** to talk about the problems or faults of someone or something
J	motivate	**10)** to do something to solve a problem
K	illness	
L	get rid of	

Reading Levels II and IV

The previous analysis was focused on listening tests. This part concentrates on the analysis of reading tests for Levels II and IV. As explained before, the tests included in the analysis were: In level II: 201510, 201530, 201610, 201630 and 201710; level IV: 201530, 201610, 201630 and 201710) with the assistance of the checklist. "Yes", "No", and "N/A" were tabulated per question in all the exams prior and after Test Specs. These questions were also divided into three categories (Validity, Text language level appropriacy, and Test items quality). The impact of Test Specs was initially measured by illustrating a graphic per question where "Yes", "No", and "N/A" were visibly identified in the listening tests of level II and IV mentioned above.

Validity

1. Does the test accurately measure what it intends to measure?

Before Specs

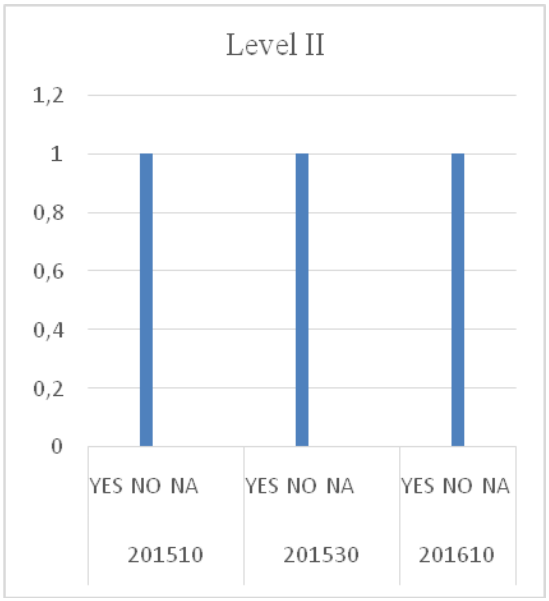


Figure 73

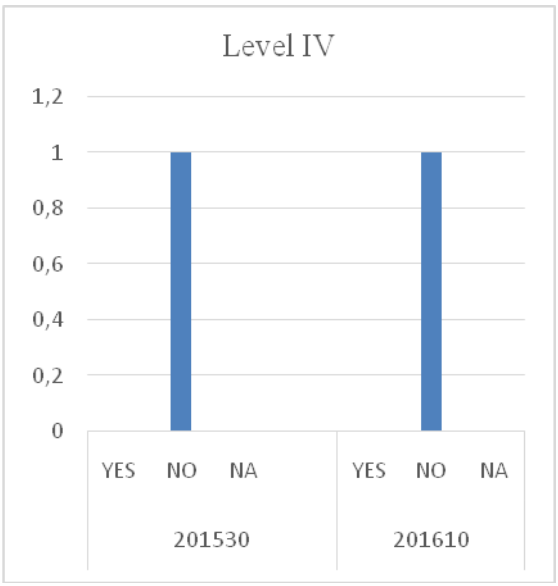


Figure 74

Figures 73 and 74 show whether tests, before test specs, accurately measure what they intend to measure. As observed in the figures, *201510, 201530*, and *201610* tests of II level do not assess what they are expected to. Likewise, *201530 and 201610* tests of level IV do not evaluate the outcomes of the levels. In other words, none of the tests comply with this validity principle since they are not measuring what they intend to measure. For example, in the 2015 level II test included these items; however, some questions asked in the test did not correspond to any of the outcomes set for the course.

After Specs

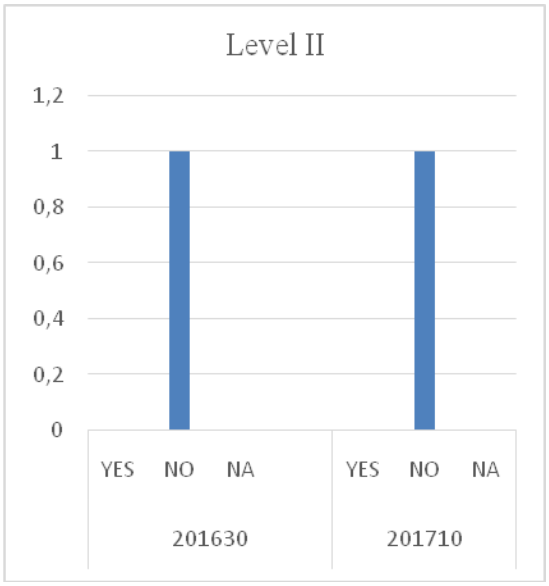


Figure 75

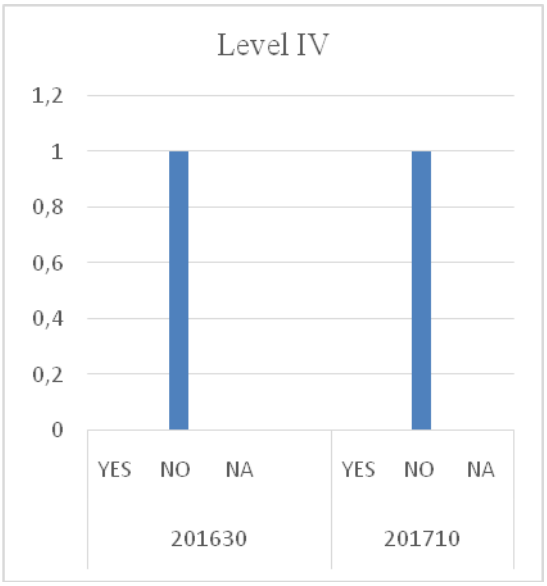


Figure 76

Figures 75 and 76 illustrate whether tests, after test specs, accurately measure what they intend to measure or not. As shown in the figures, *201630 and 20710* tests of level II assess different outcomes of the level, and not necessarily what they are expected to assess. Similarly, *201630 and 201710* tests of level IV do not evaluate the intended outcomes of level IV. This

means that all tests in both level II and IV are not valid enough due to the fact that they are not measuring what they aim to.

Sample of test and skills assessed

DO NOT WRITE ON THIS PAPER---TEST BOOKLET---

Which of these ideas are mentioned in the text? Put a check mark next to the 4 ideas that are discussed in the passage. You will not use all of the sentences. (4 points) .

1. _____ A person who finds inspiration in little things and makes a big business around it.
2. _____ Sometimes, investigating about your possible product is a good idea.
3. _____ Just because someone doesn't use it, does not mean it is outdated or cannot be used again.
4. _____ When all your options are gone, make sure you go to another place to find inspiration.
5. _____ Even if you don't succeed at first, it is always a good idea to try another time.
6. _____ The need to find a better income and a better working schedule are big motivations.

B.READING FOR DETAILS

Paraphrasing: Circle the sentence, a or b, that expresses more closely the idea of the original sentence (3 points).

7. Brian said balancing work and studies was challenging, but he did not forget to pay attention in class.

a. Even though his job was time consuming, he still managed to have a good academic record in school.
b. His job was very demanding and his studies were very difficult, so he had a challenging time in school.

8. However, some students have great ideas that simply cannot wait until graduation day.

a. Some people go to work without studies.
b. Some people start their business while in school.

9. "But I'm learning more now than I ever have in the classroom."

a. I have never discovered so much in a regular lesson.
b. I never learned anything in my normal lessons.

C. INFERRING MEANING FROM CONTEXT

Match each word with its definition. The words in underline bold are from the passage. (8 points)

10. Resource	**d.** To get bigger
11. Entrepreneur	**e.** The land and buildings of a university or college
12. Retail	**f.** Concerning the sale of things to people in stores
13. Concept	**g** Motivate, inspire
14. Stationery	**h** An idea about how something is or should be done
15. Expand	
16. Spur	
17. Campus	

2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?

Before Specs

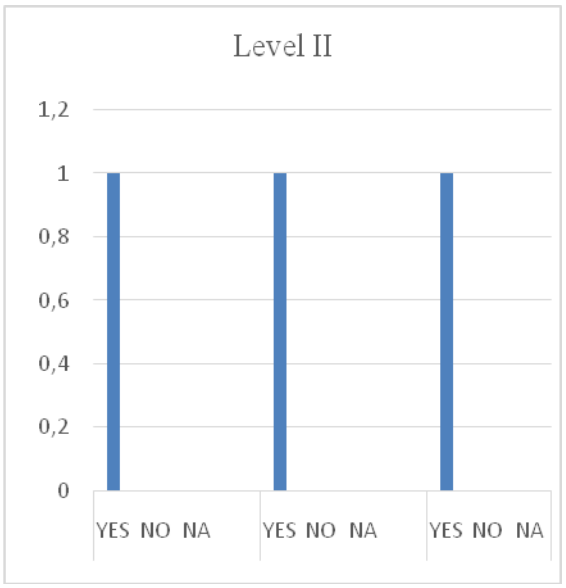


Figure 77

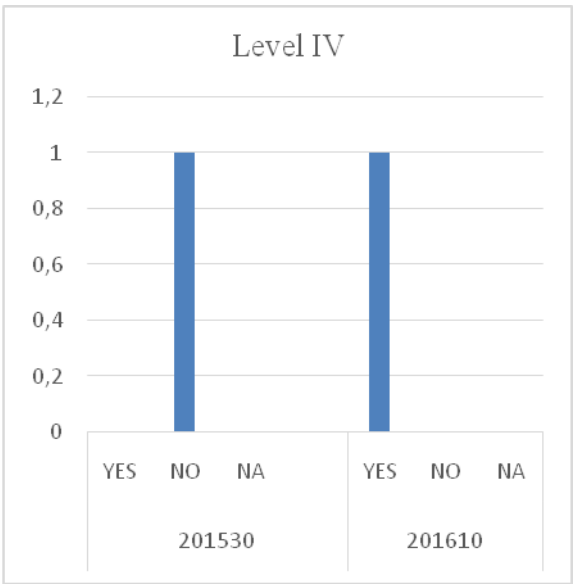


Figure 78

Figures 77 and 78 illustrate the reasonability of the number of questions asked in the tests considering a time frame before Specs. Evidently, *201510, 201530* and *201610* tests of level II IV and *201610* of level do have reasonable number of questions that can be completed by students within the expected time frame. However, as observed in figure 78, 201530 test of level IV do not possess reasonable amount of questions concerning a time frame. That is to say, the number of questions in most of level II and IV tests are reasonable, and students can actually complete them within a time frame. Even though Assessment Handbook established a fifty minute time frame for students to answer, these guidelines did not specify the approximate number of questions tests should have.

After Specs

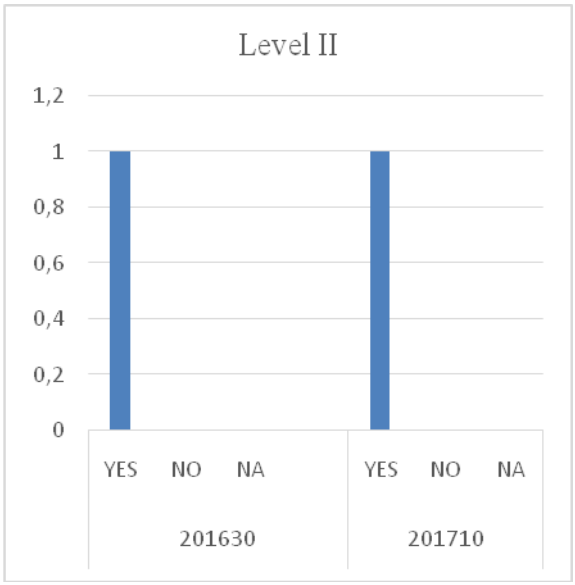


Figure 79

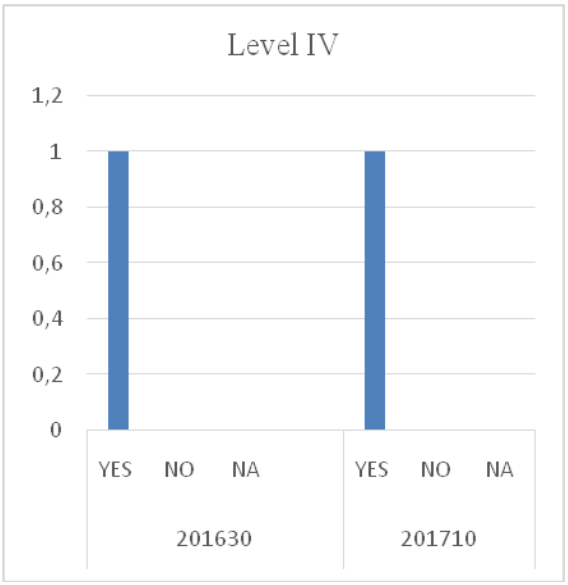


Figure 80

Figures 79 and 80 present the reasonability of the number of questions asked in the tests considering a time frame after Specs. Clearly, *201630* and *201710* tests of level II have reasonable number of questions that can be completed by students within the expected time frame. Likewise, *201630* and *201710* tests of level IV also have reasonable amount of questions concerning the time frame. In this way, the number of questions in all tests of both level II and level IV are reasonable and students can complete them within the time frame. Even though Test Specs established a fifty minute time frame for students to answer, these guidelines did not specify the approximate number of questions tests should have.

3. Are the items well distributed to make the test valid?

Before Specs

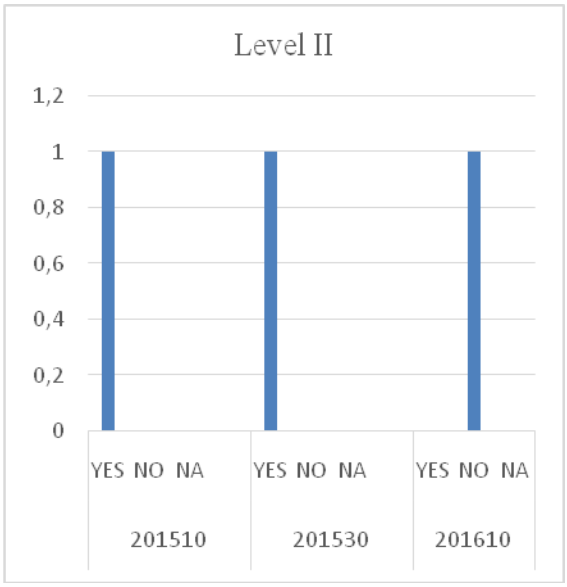


Figure 81

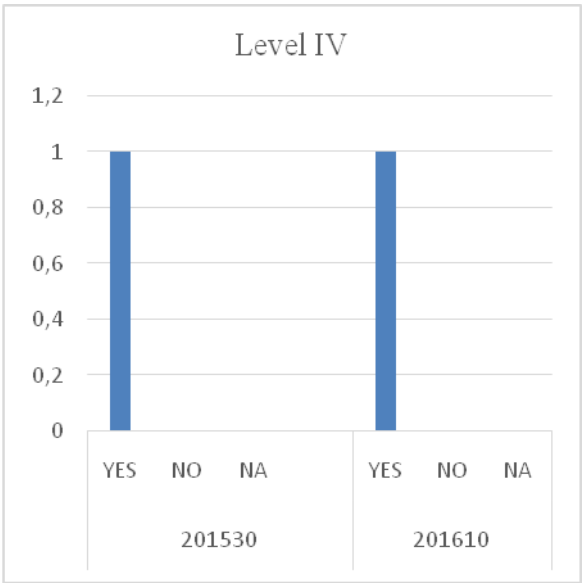


Figure 82

Figures 81 and 82 present the distribution of the items in tests to make them valid before Specs. This means that each section has similar amount of questions and points. As shown in the figures, *201510, 201530* tests of level II and *201530* and *201610* tests of level IV have well distributed items that make tests valid. On the contrary, *201610* test of level II does not have items that are well distributed. As it can be seen, most of tests in both levels II and IV comply the principle of validity in terms of the distributions of the items. According to the guidelines used to design tests, sections should all be of similar length and structure.

After Specs

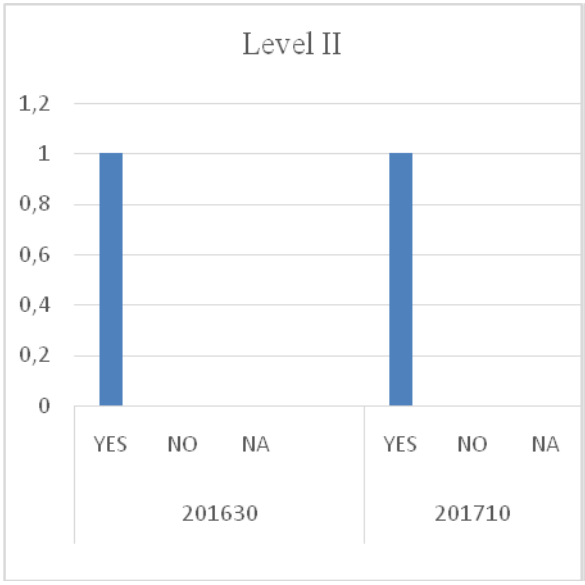


Figure 83

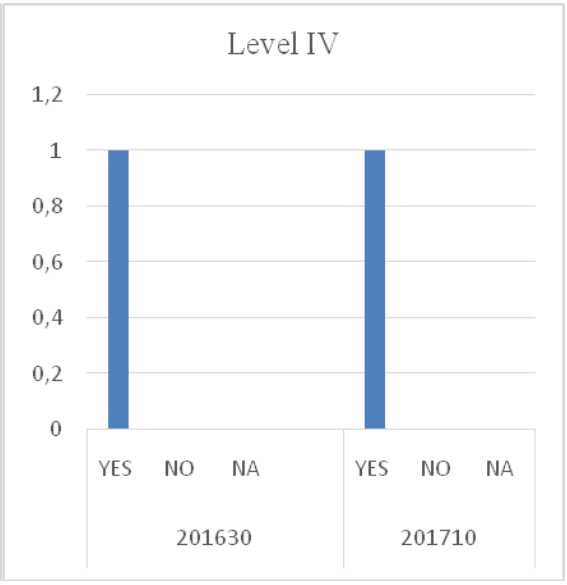


Figure 84

Figures 83 and 84 present the distribution of the items in tests to make them valid after Specs. As illustrated in these figures, *201630* and *201710* tests of level II do have well distributed items that make tests valid. *201630* and *201710* test of level IV similarly distribute the items properly. Thus, all the tests in both levels II and IV after specs fulfill the principle of validity in terms of the distributions of the items. According to the Test Specs used to design tests, sections should all be of similar length and structure.

4. Does test construction follow specific guidelines?

Before Specs

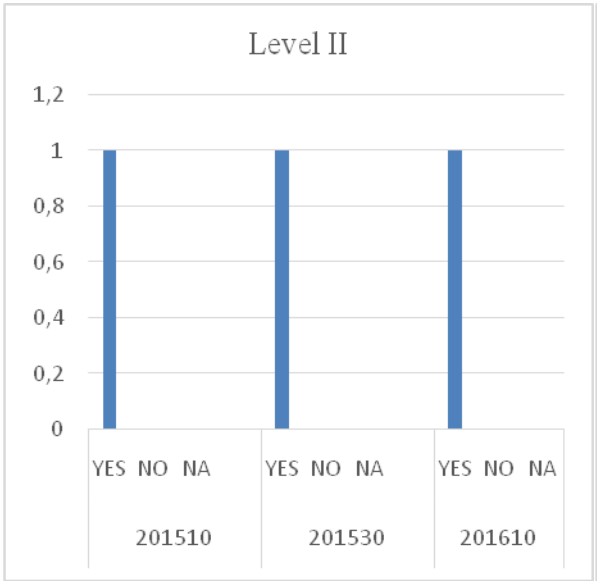


Figure 85

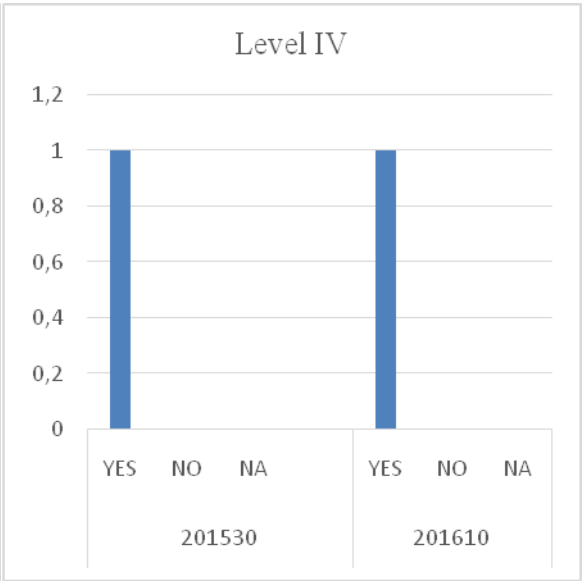


Figure 86

Figures 85 and 86 suggest whether tests followed specific guidelines or not. As presented in the figures, *201510*, *201530*, *201610* tests of level II and *201530/210610* tests of level IV were constructed following a specific guideline called Assessment Handbook. So, according to the analysis done for this study, all the tests were designed following detailed instructions. (Assessment Handbook)

After Specs

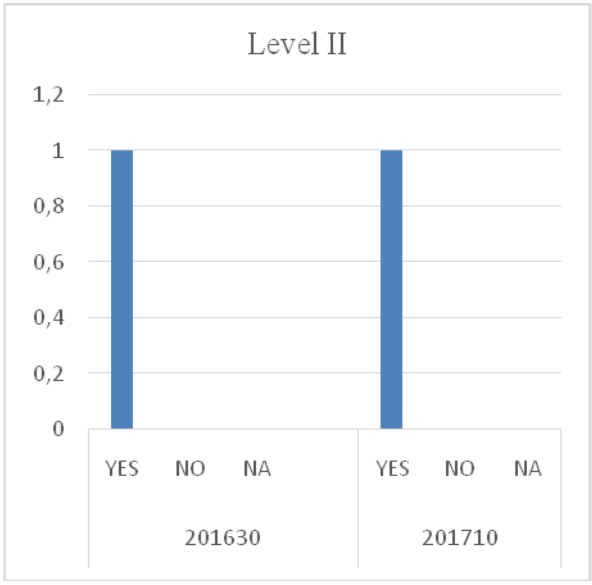


Figure 87

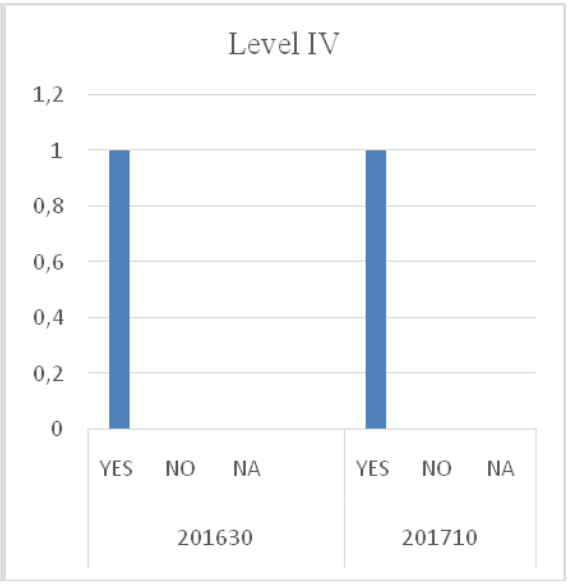


Figure 88

Figures 87 and 88 report whether tests followed specific guidelines or not. As shown in the figures, *201630* and *201710* tests of level II and *201630* and *201710* tests of level IV were constructed following a specific guideline called Test Specifications. Hence, according to the analysis done for this study, all the tests were designed following detailed instructions of tests. (Test Specs)

Texts Language Level Appropriacy

The following questions deal with the analysis of language level appropriacy. As explained in the listening section, it focuses on real-world language use, texts appropriacy, contextualized items, unambiguous items/tasks, and whether the test was developed with specific purpose and a particular group of test takers. These questions are then followed by test item quality questions.

1. Is the language in the test representative of real-world language use?

Before Specs

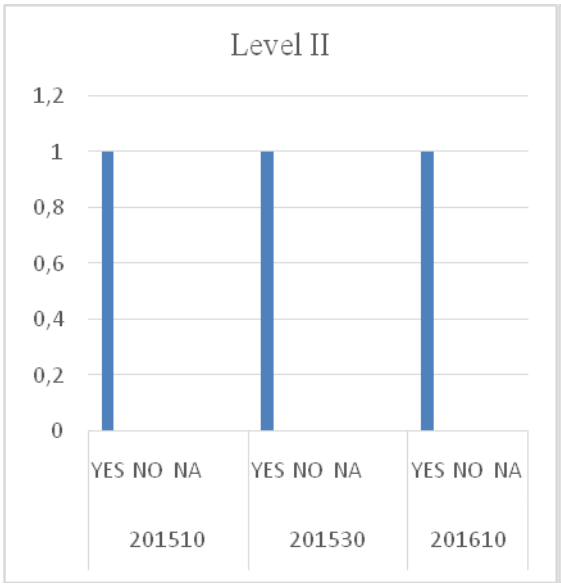


Figure 89

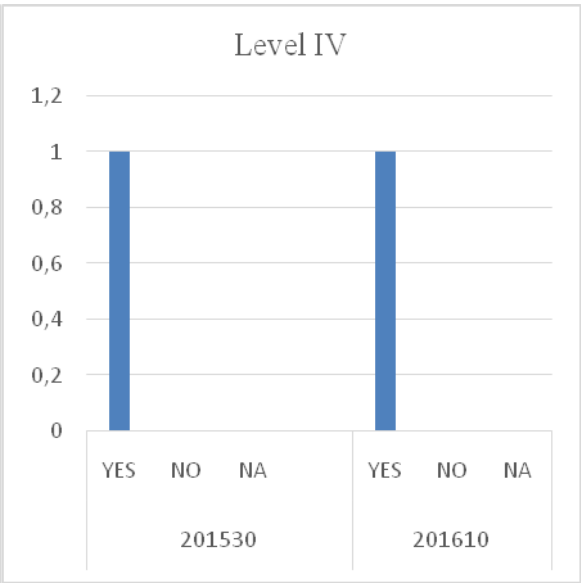


Figure 90

Figures 89 and 90 contain information concerning the language used in tests. As stated in the figures, *201510*, *201530*, and *201610* tests of level II *201530*, *201610* tests of level IV do use language that simulates real-world use. That is to say, the language used in all the tests examined in this project reflected in some way the language people use in daily life. Assessment

Handbook states that language should reflect actual use. Some topics are: *Food*, *Urban farmer and Breaking ups.*

Part Four: Reading Two - Read about farms inside of cities. *(15 minutes)*

The Urban Farmer

Instead of growing fruits and vegetables in fields, miles away from cities, grow them in the city. A school in New York City has constructed a small farm on part of its playground, and the organic produce is sold during the summer at local farm stands. One company is testing indoor farming. The whole farm is expected to grow 15 times more spinach than a traditional farm, but use only 5 percent of the amount of water. There are many experiments like these happening in cities. Michael Dickson, a professor of environmental sciences, believes that one day skyscrapers, or tall buildings, will be farms, providing enough food for entire cities. He describes the advantages of indoor farming, "You can control nothing outdoors, and you can control everything indoors." Crops grown in a controlled environment will not be threatened by nature. Also, the costs will be less because no herbicides or pesticides would be used, and the food would not need to be transported great distances. Unfortunately, urban farms do not yet produce enough food for people in cities.

A. Read for Main Ideas & Details: Read each statement. Decide if it is true or false. Write ***T*** (true) or ***F*** (false) next to it. If the statement is false, change a word or phrase to make it true. *(4 points; 1 point per question)*

_____ 1) The number of small urban farms is growing.

After Specs

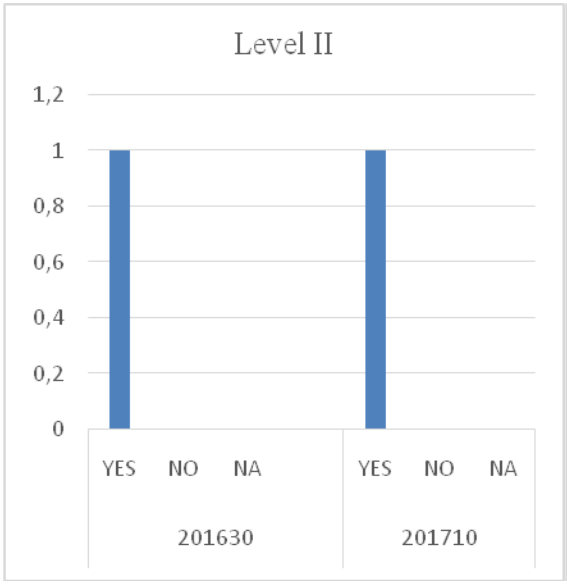


Figure 91

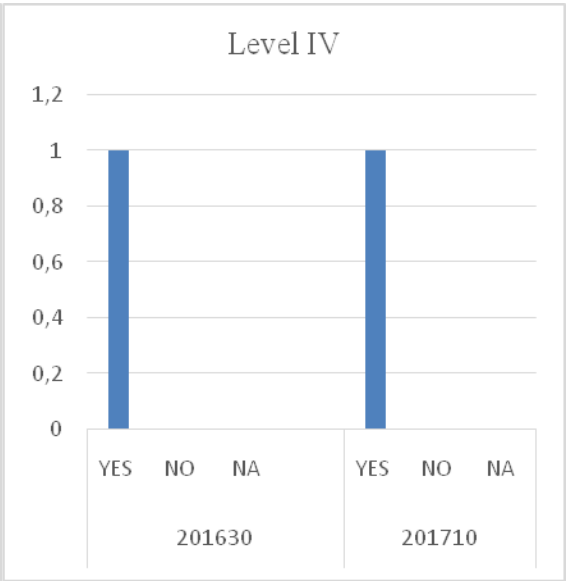


Figure 92

Figures 91 and 92 show information about the language used in tests. As it is evident in the figures, *201630* and *201710* tests of level II *201630, 201710* tests of level IV do use language that reflects the language people use in real life. In other words, the language used in all the tests analyzed for this project simulates real-world use. Test Specs state that language should reflect actual use. Some topics are: *Climate change controversy and College startups.*

READING TWO: (10 points)

Read the passage about climate change. Use the information you read to answer the questions below.

CLIMATE CHANGE CONTROVERSY

18._____. In the past it has altered as a result of natural causes. Nowadays, however, the term climate change is generally used when referring to changes in our climate which have been identified since the early part of the 1900's. The changes we have seen over recent years and those which are predicted over the next 80 years are thought to be mainly as a result of human behaviour rather than due to natural changes in the atmosphere. The greenhouse effect is very important when we talk about climate change as it relates to the gases keeping the Earth warm. It is the extra greenhouse gases which humans have released that are thought to pose the strongest threat.

19._____. The report "Are we putting our fish in hot water?" describes how climate change is causing temperatures to rise in rivers, lakes and seas. This means less food and oxygen for marine life, damaged fish growth and fewer reproduction. The report says that temperate fish such as salmon, catfish and sturgeon cannot produce eggs if winter temperatures do not drop below a specific level. Warmer water also means fish could extensively migrate to cooler areas, where the temperature is similar to their normal habitat. This could impact on many species' ability to survive. Some species will become extinct if the water temperature increases by a degree or two.

20._____. WWF director Andrew Lee said: "Climate change increases the pressure on fish populations that are already strained to the limit by over-fishing in the marine environment. We must act urgently to reduce both carbon dioxide emissions and fishing pressures to protect fish populations as they are one of the world's most valuable biological, nutritional and economic assets." Forty percent of the world's people are reliant on fish for basic sustenance and a main source of protein. Dr Richard Dixon, director of WWF Scotland spoke ahead of next week's UN Climate Change Conference in Montreal. He said: "If we fail to secure deeper reductions in greenhouse gas emissions we will increase the pressures on fish and the billions of people that depend on them."

21._____. The list of things we need to think about which will be affected by climate change is endless. In this section we give you a few examples of how we will need to change the way we live in order to cope with changes to our climate. The regular use of renewable energy is becoming increasingly popular. Have a look at the possibilities for alternative energy sources, including solar power, wind power, geothermal, water power and even nuclear energy….-"What else can you do to help adapt to climate change and what can you do to help slow it down? There are many things we can all do at home…"

2. Are the readings level appropriate?

Before Specs

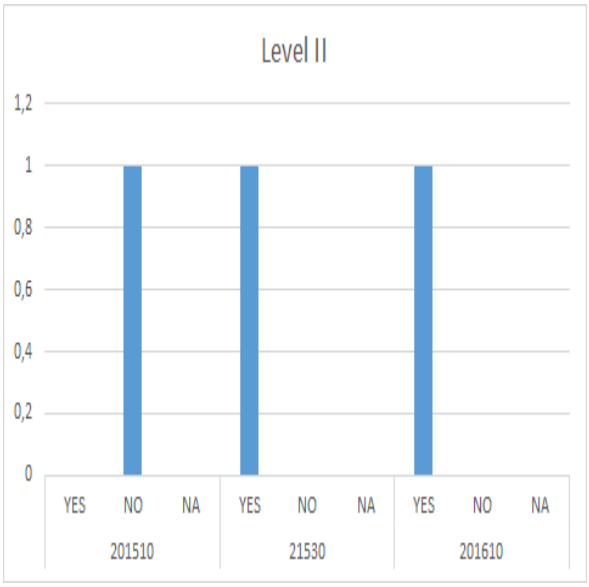


Figure 93

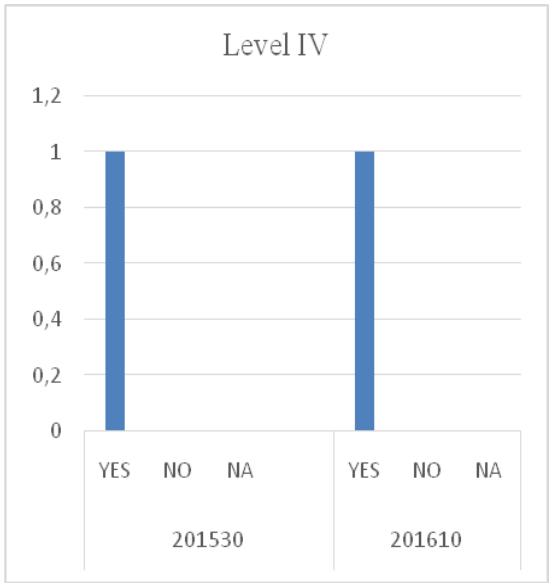
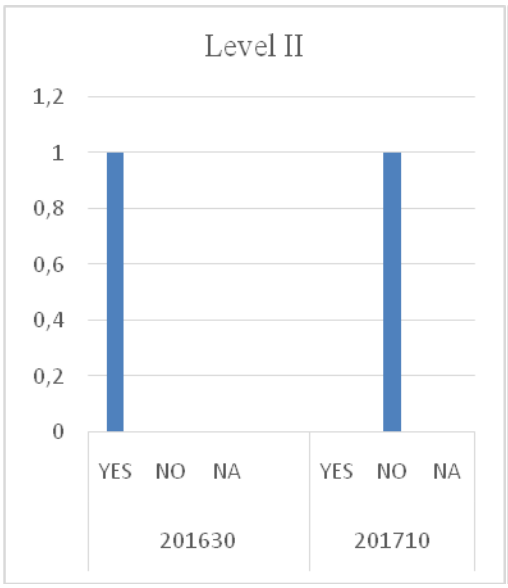


Figure 94

Figures 93 and 94 present the level appropriateness of the readings used in tests. It is clear that in *201510* test, the level of the readings is not appropriate. In contrast, *201530*, *201610* tests of level II and *201530* and *201610* tests of level IV provide students with texts that are suitable for them. This means that almost all the tests studied in this project meet the needs of students with appropriate text level.

After Specs



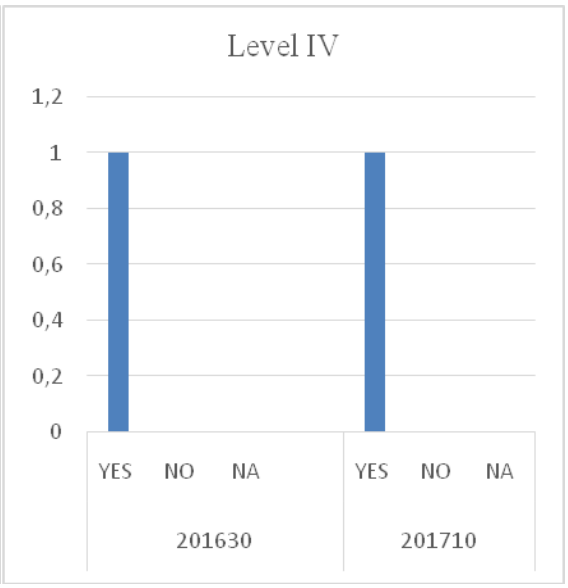


Figure 95 Figure 96

Figures 95 and 96 display the level appropriateness of the readings used in tests after Specs. As illustrated in the figures, *201630* test of level II and *201630* and *201710* tests of level IV contain texts that are proper for students' level. Opposite to this, in the *201510* test of level II, the readings are not level appropriate. To be more precise, almost all the examined tests include readings that are within the appropriate text level.

3. Does the test have items that are contextualized rather than isolated?

Before Specs

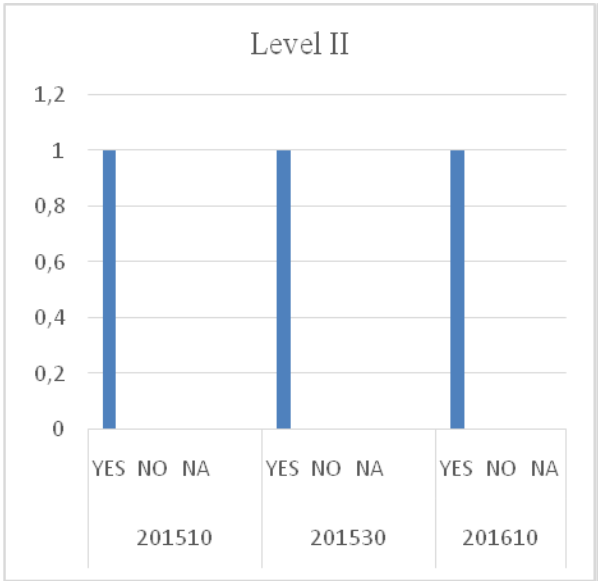


Figure 97

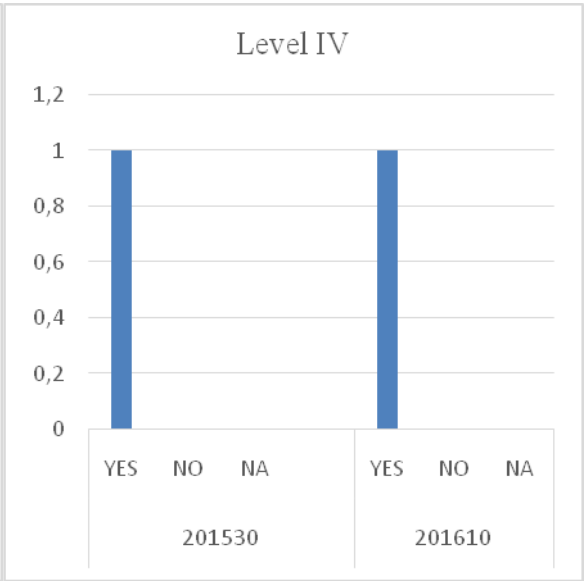


Figure 98

Figures 97 and 98 show if items in the tests are contextualized rather than isolated. All the tests (*201510*, *201530*, and *201610* tests of level II and *201530* and *201610* tests of level IV) have items that are contextualized with each other and with the outcomes of each level. Accordingly, most of the questions asked in the tests analyzed are coherent with the objectives of the levels.

After Specs

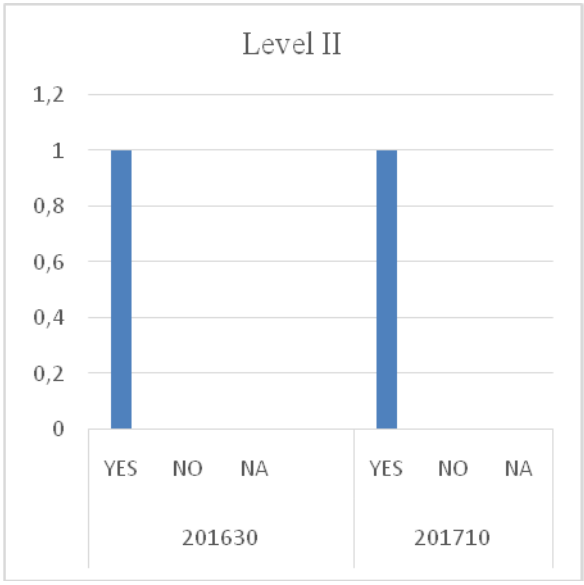


Figure 99

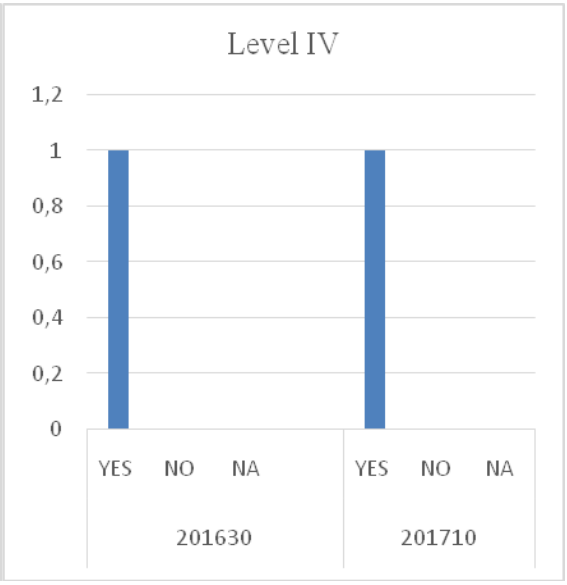


Figure 100

Figures 99 and 100 read if items in the tests are contextualized rather than isolated. *201630*, *201710*, tests of level II and *201630* and *201710* tests of level IV include questions that are contextualized with the objectives of the levels. Thus, all the items in the tests analyzed are consistent with each other and with the outcomes of each level.

4. Does the test contain items/tasks that are unambiguous to the test taker?

Before Specs

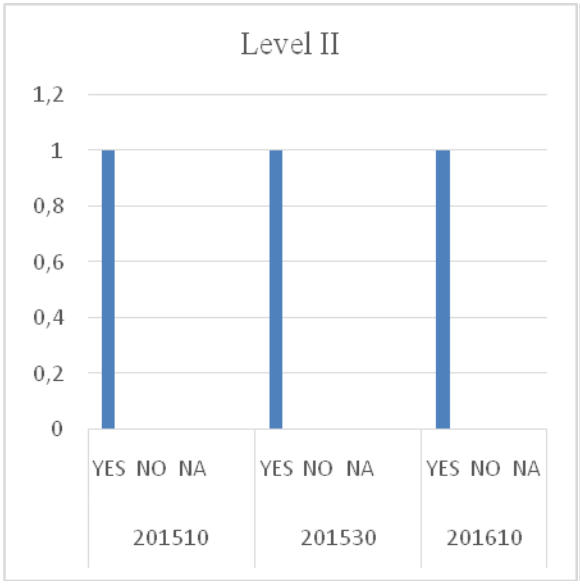


Figure 101

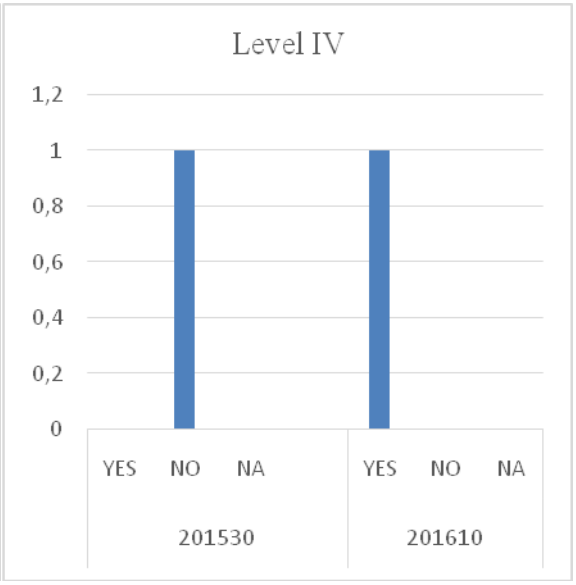


Figure 102

Figures 101 and 102 present whether the items and/or tasks in the tests are unambiguous to the test takers. As explained before, this means that the items and/or tasks are not open to more than one interpretation. As illustrated in the figures, *201510*, *201530*, and *201610* test of level II and *201610* tests of level IV have unambiguous items and/or tasks. However, *201530* test of level IV do not contain items and/or tasks easy to understand for the test-takers. So, most of the tests evaluated in this study have clear questions and exercises for students.

After Specs

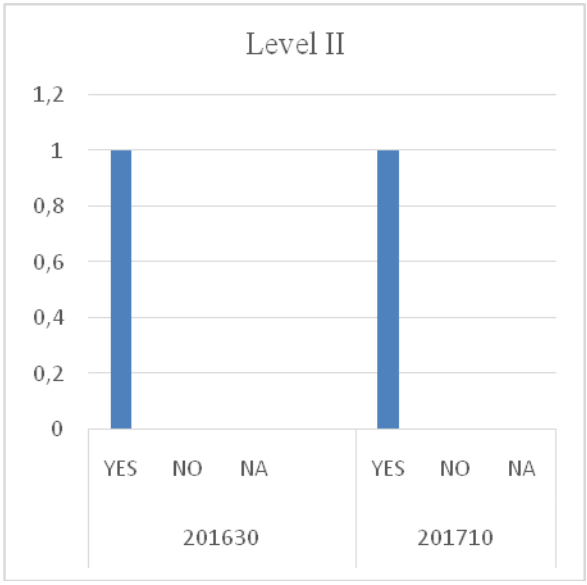


Figure 103

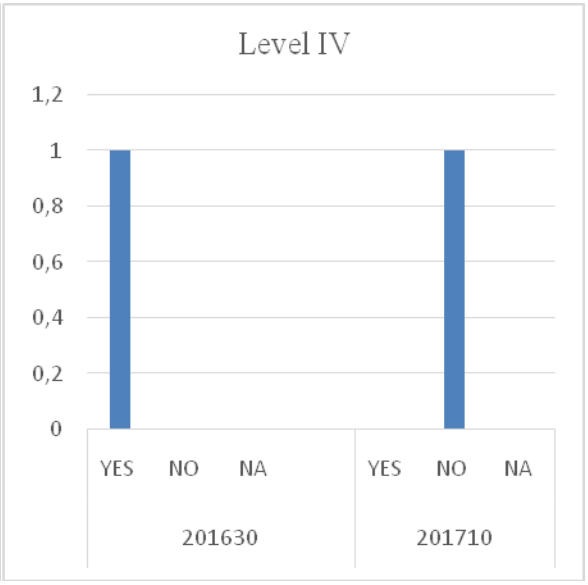


Figure 104

Figures 103 and 104 suggest information regarding the ambiguity of the items and/or tasks in the tests. As marked in the figures, *201630* and *201710* tests of level II and *201630* test of level IV include clear items and/or tasks. However, *201710* test of level IV have some problematic items and/or tasks for students. So, most of the tests examined in this study have unambiguous questions and exercises for the test takers.

5. Is the test developed with a specific purpose and a particular group of test-takers?

Before Specs

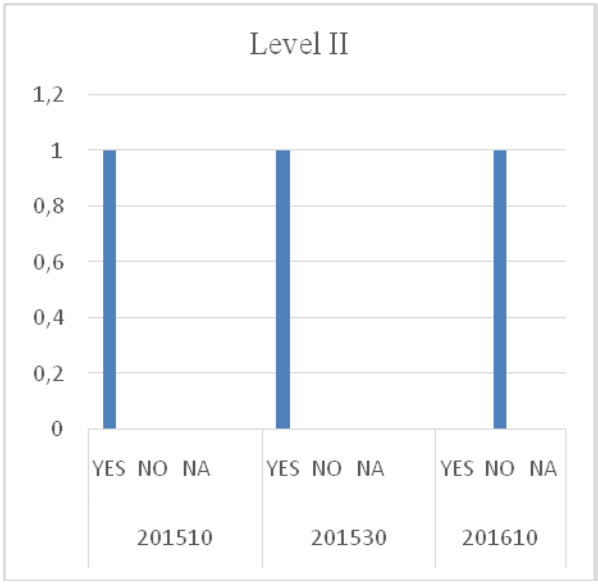


Figure 105

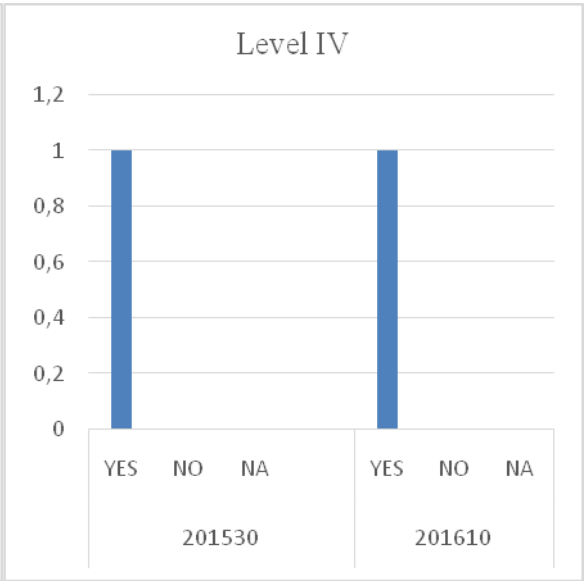


Figure 106

Figures 105 and 106 read information regarding the specific purpose and the particular groups of test-takers with which the tests are developed. It is quite visible that *201510* and *201530* tests of level II and *201530* and *201610* tests of level IV do have a specific purpose and a particular group of test-takers. On the other hand, the *201610* test of level II does not consider all these important aspects. Hence, almost every test analyzed in this study shows to have been constructed taking into account the groups of students taking the tests, and the outcomes of the levels, since these tests were conceived considering the level in which students should be. (See sample)

Sample

Universidad del Norte – Instituto de Idiomas
Level Four: **Reading Assessment ver A**

Name: _____
Date: _____

Reading One: Read the entire story. Use the information within to choose the correct answers.

Breaking Up

[1] It's never easy when a marriage or significant relationship ends. Whatever the reason for the breakup—and whether you wanted it or not—the end of a relationship can turn your whole world upside down and trigger all sorts of painful feelings. But there are plenty of things you can do to get through this difficult time and move on. You can even learn from the experience and grow into a stronger, wiser person.

[2] Why do breakups hurt so much, even when the relationship is no longer good? A divorce or breakup is painful because it represents the loss, not just of the relationship, but also of shared dreams and commitments. Romantic relationships begin on a high note of excitement and hope for the future. When these relationships fail, we experience profound disappointment, stress, and sadness.

[3] A breakup or divorce launches us into unknown territory. Everything is disrupted: your routine and responsibilities, your home, your relationships with extended family and friends, and even your identity. A breakup brings uncertainty about the future. What will life be like without your partner? Will you find someone else? Will you be forever alone? These unknowns often seem worse than an unhappy relationship.

[4] Recovering from a breakup or divorce is difficult. However, it's important to know (and to keep reminding yourself) that you can and will move on. But healing takes time, so be patient with yourself.

[5] Allowing yourself to feel the pain of these losses may be scary. You may fear that your emotions will be too intense, or that you'll be stuck in a dark place forever. Just remember that grieving is essential to the healing process. The pain of grief is precisely what helps you let go of the old relationship and move on. And no matter how strong your grief, it won't last forever.

1. **Read for Main Ideas:** Check (✓) the best answer to the following questions (8 points).

The sample provides information concerning the group of test-takers and also the purpose of the exam which is to evaluate the reading outcomes of level four. The language in the test is familiar to students at this level.

After Specs

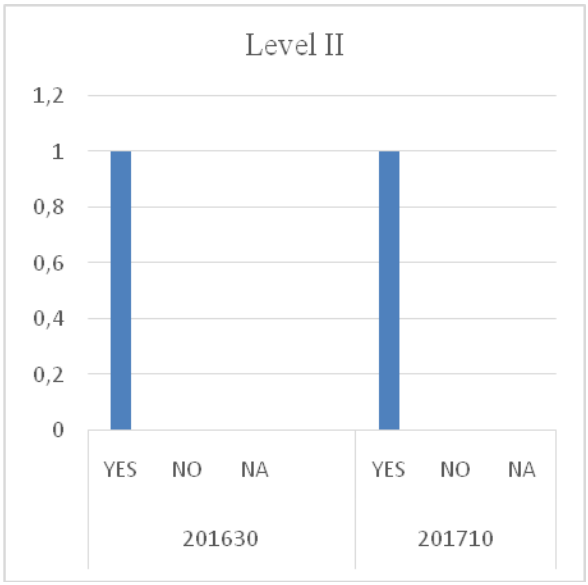


Figure 107

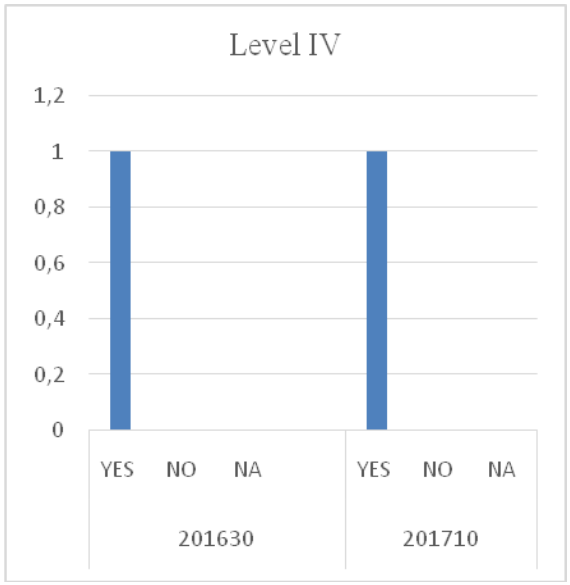
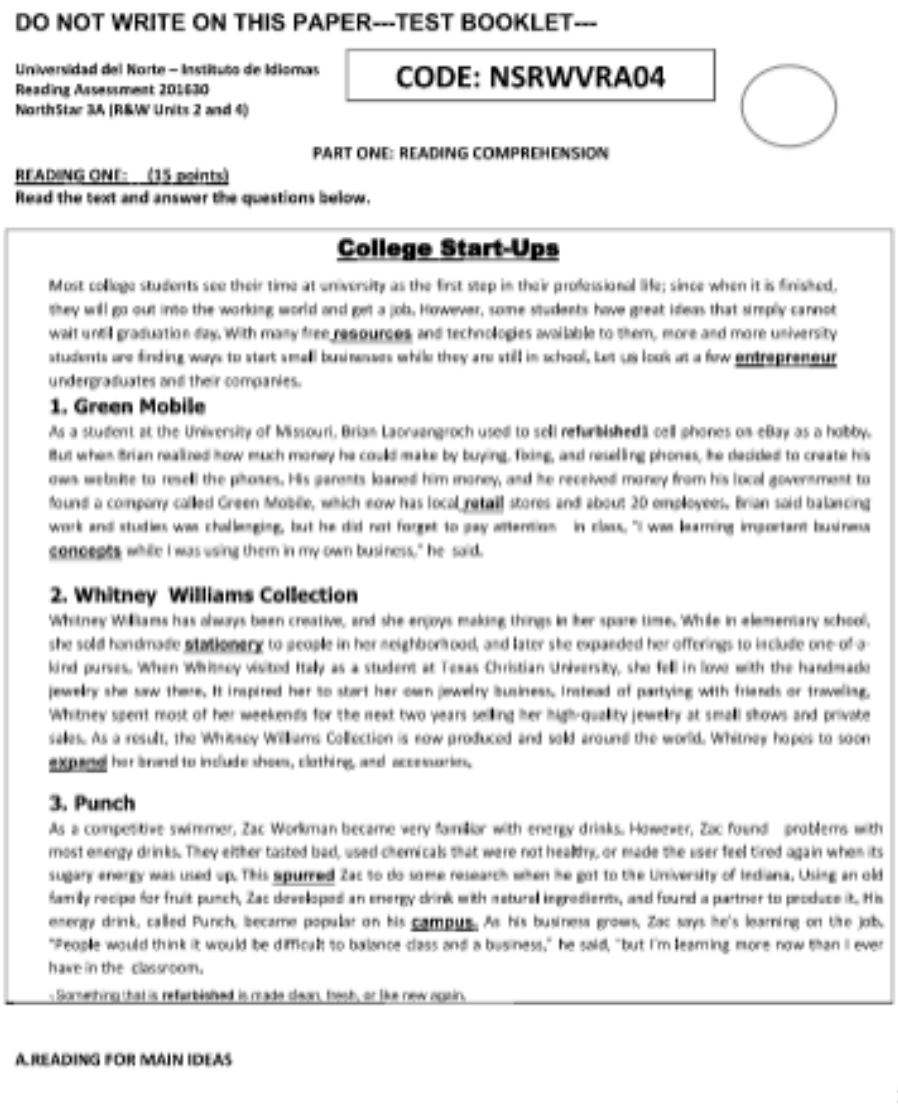


Figure 108

Figures 107 and 108 show information concerning the specific purpose and the particular groups of test-takers with which the tests are developed. *201630* and *201710* tests of level II and *201630* and *201710* tests of level IV consider a specific purpose and a particular group of test-takers. Therefore, every test examined in this study has been designed taking into account the groups of students taking the tests, and the outcomes of the levels. (See sample)

Sample



DO NOT WRITE ON THIS PAPER---TEST BOOKLET---

Universidad del Norte – Instituto de Idiomas
Reading Assessment 201630
NorthStar 3A (R&W Units 2 and 4)

CODE: NSRWVRA04

PART ONE: READING COMPREHENSION

READING ONE: (15 points)
Read the text and answer the questions below.

College Start-Ups

Most college students see their time at university as the first step in their professional life; since when it is finished, they will go out into the working world and get a job. However, some students have great ideas that simply cannot wait until graduation day. With many free **resources** and technologies available to them, more and more university students are finding ways to start small businesses while they are still in school. Let us look at a few **entrepreneur** undergraduates and their companies.

1. Green Mobile

As a student at the University of Missouri, Brian Laoruangroch used to sell **refurbished**1 cell phones on eBay as a hobby. But when Brian realized how much money he could make by buying, fixing, and reselling phones, he decided to create his own website to resell the phones. His parents loaned him money, and he received money from his local government to found a company called Green Mobile, which now has local **retail** stores and about 20 employees. Brian said balancing work and studies was challenging, but he did not forget to pay attention in class, "I was learning important business **concepts** while I was using them in my own business," he said.

2. Whitney Williams Collection

Whitney Williams has always been creative, and she enjoys making things in her spare time. While in elementary school, she sold handmade **stationery** to people in her neighborhood, and later she expanded her offerings to include one-of-a-kind purses. When Whitney visited Italy as a student at Texas Christian University, she fell in love with the handmade jewelry she saw there. It inspired her to start her own jewelry business. Instead of partying with friends or traveling, Whitney spent most of her weekends for the next two years selling her high-quality jewelry at small shows and private sales. As a result, the Whitney Williams Collection is now produced and sold around the world. Whitney hopes to soon **expand** her brand to include shoes, clothing, and accessories.

3. Punch

As a competitive swimmer, Zac Workman became very familiar with energy drinks. However, Zac found problems with most energy drinks. They either tasted bad, used chemicals that were not healthy, or made the user feel tired again when its sugary energy was used up. This **spurred** Zac to do some research when he got to the University of Indiana. Using an old family recipe for fruit punch, Zac developed an energy drink with natural ingredients, and found a partner to produce it. His energy drink, called Punch, became popular on his **campus.** As his business grows, Zac says he's learning on the job. "People would think it would be difficult to balance class and a business," he said, "but I'm learning more now than I ever have in the classroom.

1 Something that is **refurbished** is made clean, fresh, or like new again.

A.READING FOR MAIN IDEAS

1

The sample provides information concerning the group of test-takers and also the purpose of the exam which is to evaluate the reading outcomes of level four.

Test Items Quality

The questions below explore the test item quality.

1. Is the stem clear and precise (It clearly indicates the kind of answers that students need to give)?

Before Specs

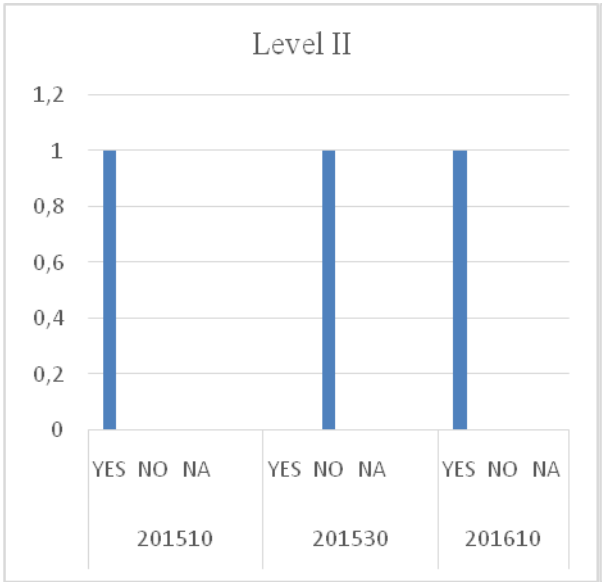


Figure 109

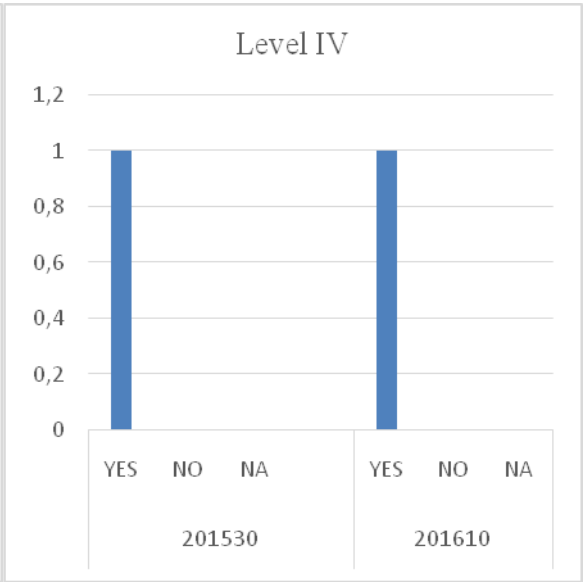
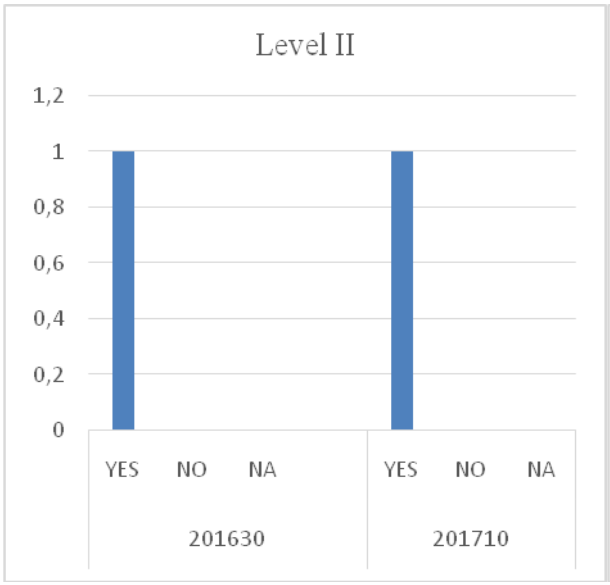


Figure 110

Figures 109 and 110 show the clarity of the stem whether or not it clearly indicates what students have to do in the test. As can be seen, while *201510* and *201610* tests of level II had stems that were clear and precise, *201530* test of level II did not clearly indicate what students have to do. In addition, *201530* and *201610* tests of level IV also comply with the clarity and precision of the stem even before Specs. Hence, most of the tests studied for this study provide students with accurate instructions so that they can efficiently complete the assessment.

After Specs



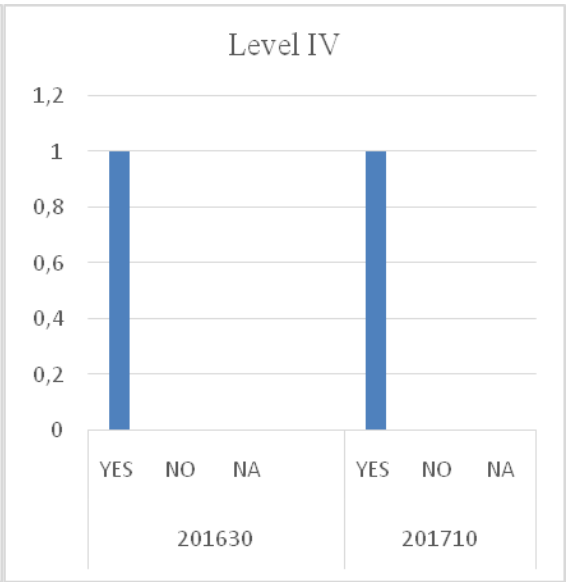


Figure 111 Figure 112

Figures 111 and 112 describe the clarity of the stems after Specs. As illustrated in the figures, *201630* and *201710* tests of level II and 201630 and 201710 tests of level IV have stems that are clear and precise. This means that they clearly indicate what students have to do in their tests. Therefore, all the tests in both level II and IV comply with quality of tests since these are giving students the proper instructions and questions.

2. Is each option clearly identified as the answer to the question asked?

Before Specs

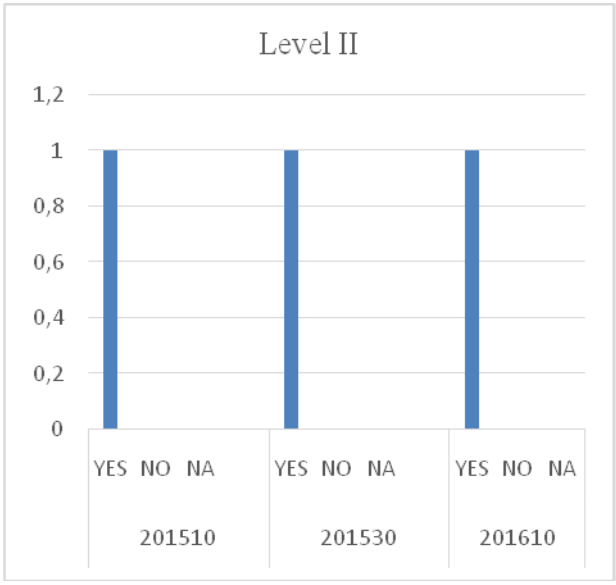


Figure 113

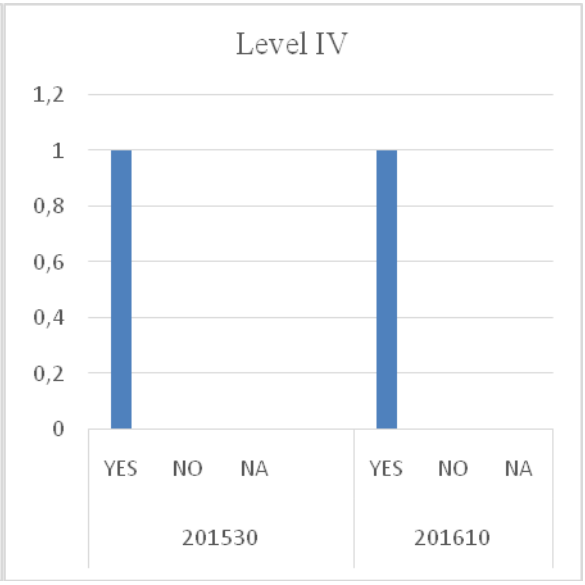


Figure 114

Figure 113 and 114 illustrate if each option is clearly identified as the answer to the question asked before Specs. Evidently, 201510 tests of level II and 201530 and 201610 tests of both levels II and IV do have understandable options or answers to the questions asked. They are clearly identified as the choices of the questions to be selected in the tests.

After Specs

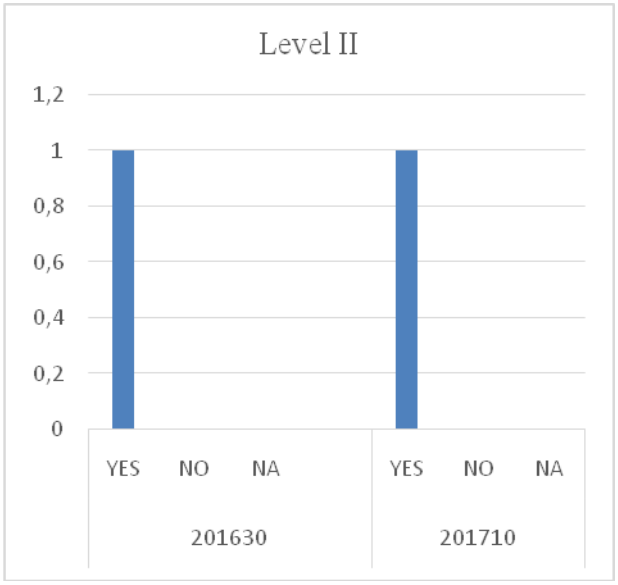


Figure 115

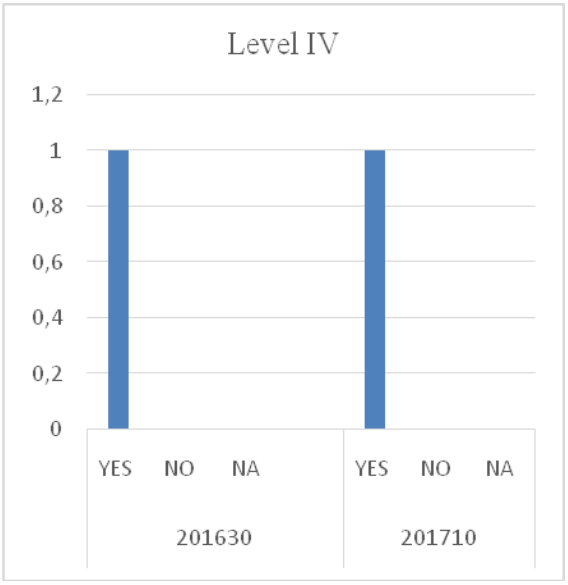
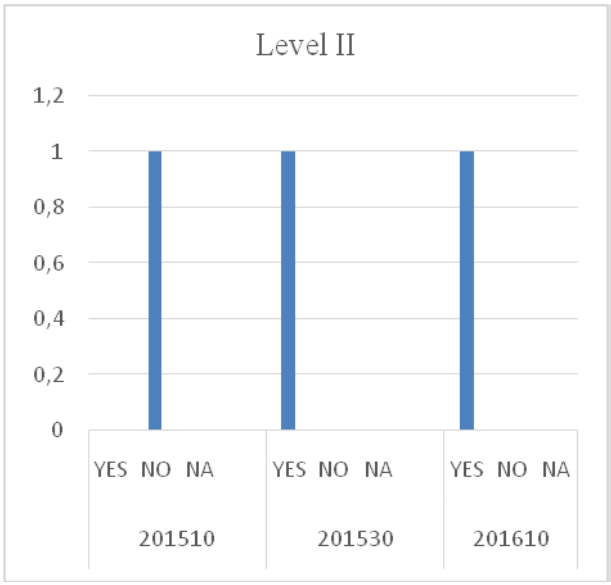


Figure 116

Figure 115 and 116 show if each option is clearly identified as the answer to the question asked after Specs. Visibly, all 201630 and 201710 tests of both level II and IV possess clear options or answers to the questions asked. They are clearly recognized as the choices of the questions to be selected in the tests.

3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)

Before Specs



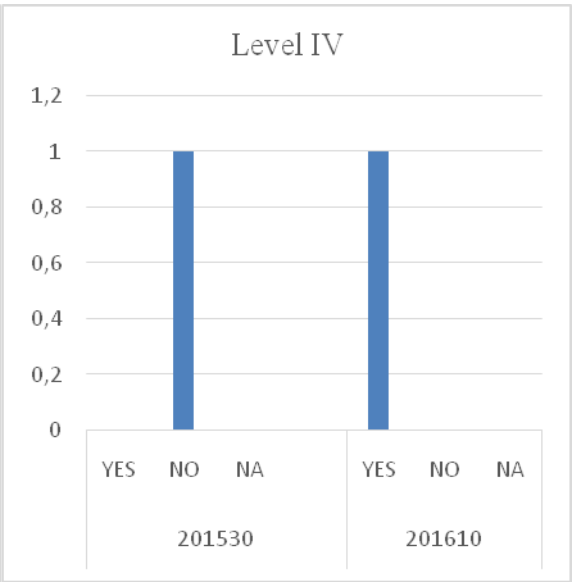


Figure 117 Figure 118

Figure 117 and figure 118 read the text dependency that answers of questions have before Specs. *201510, 201530* and *201610* tests of level II and 201610 test of level IV reveal that answers to each questions are text dependent because students really need to read the text if they want to spot the right answer. In contrast to those, 201530 test of level IV does not have text dependent answers to questions since they can actually be answered by students' prior knowledge. Likewise, most of the tests in both levels II and IV have text dependent answers that require reading texts for students to answer them. (See sample)

Sample

3) In paragraph 2, how is comfort food different today than in the past?
   a) People nowadays have unhealthy diets
   b) People would rather eat junk food than comfort food
   c) Comfort food now often means a different kind of food than it did
   d) Comfort food is enjoyed internationally

4) In paragraph 2, people who eat unhealthily _____.
   a) can gain weight quickly
   b) need to be more active than those who eat well
   c) are less stressed than others
   d) eat to make themselves feel happier

5) Fresh and healthy foods _____.
   a) are the best things to eat on a diet
   b) can affect how we think about food
   c) are more comforting that fatty foods
   d) have a range of benefits

6) What effect of eating unhealthily is presented in paragraph 4?
   a) having mixed feelings of satisfaction and guilt
   b) an increase in overweight people
   c) less balanced diets than in the past
   d) a population of people who only eat comfort food

After Specs

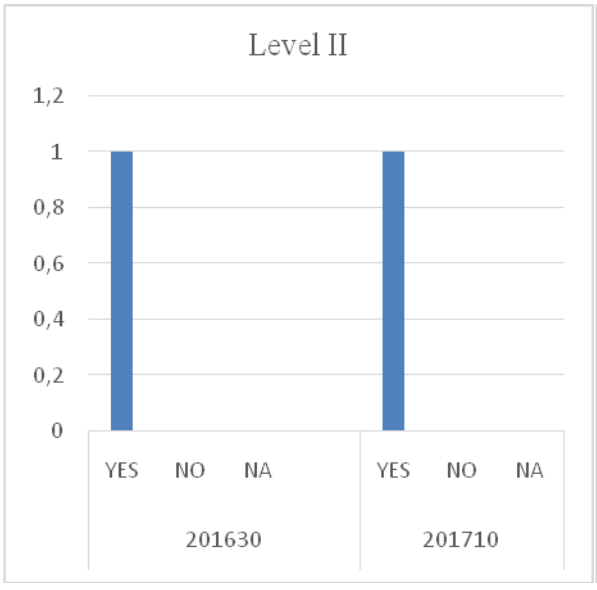


Figure 119

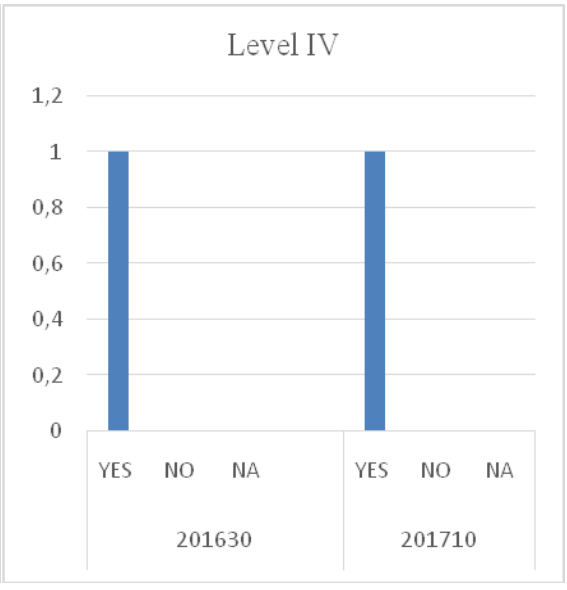


Figure 120

Figure 119 and figure 120 display the text dependency that answers of questions have After Specs. 201630 and 201710 in both levels II and IV have text dependent answers to questions. For this reason, all the tests in both levels have text dependent answers that require reading texts for students to answer the questions of assessment. (See sample)

Sample

DO NOT WRITE ON THIS PAPER---TEST BOOKLET---

Which of these ideas are mentioned in the text? Put a check mark next to the 4 ideas that are discussed in the passage. You will not use all of the sentences. (4 points) .

1. _____ A person who finds inspiration in little things and makes a big business around it.
2. _____ Sometimes, investigating about your possible product is a good idea.
3. _____ Just because someone doesn't use it, does not mean it is outdated or cannot be used again.
4. _____ When all your options are gone, make sure you go to another place to find inspiration.
5. _____ Even if you don't succeed at first, it is always a good idea to try another time.
6. _____ The need to find a better income and a better working schedule are big motivations.

4. Is the answer to each question text dependent (it does not depend on other stems and keys)

Before Specs

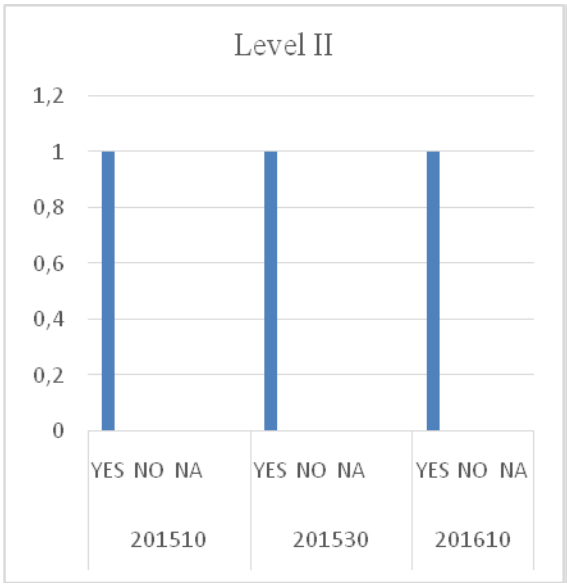


Figure 121

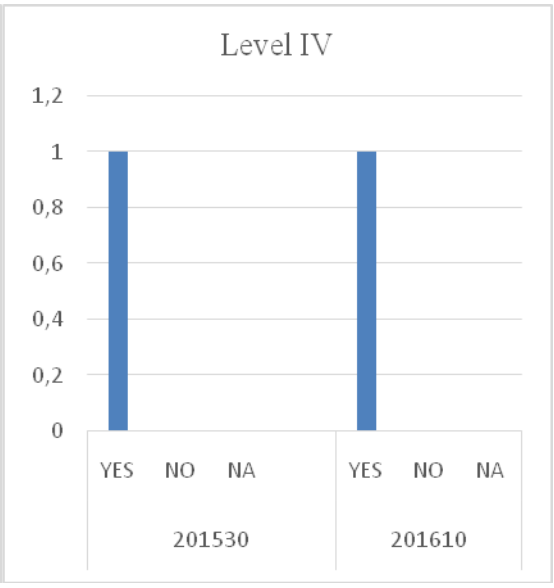


Figure 122

Figure 121 and 122 present the text dependency that answers to questions have before specs. 201510 test of level II and 201530 and 201610 tests of both levels II and IV have text dependent answers to each question asked. That is to say that all answers are text dependent because they do not depend on other stems and keys but on the text itself for students to complete the assessment. (See sample)

Sample

3) In paragraph 2, how is comfort food different today than in the past?
   a) People nowadays have unhealthy diets
   b) People would rather eat junk food than comfort food
   c) Comfort food now often means a different kind of food than it did
   d) Comfort food is enjoyed internationally

4) In paragraph 2, people who eat unhealthily _____.
   a) can gain weight quickly
   b) need to be more active than those who eat well
   c) are less stressed than others
   d) eat to make themselves feel happier

5) Fresh and healthy foods _____.
   a) are the best things to eat on a diet
   b) can affect how we think about food
   c) are more comforting that fatty foods
   d) have a range of benefits

6) What effect of eating unhealthily is presented in paragraph 4?
   a) having mixed feelings of satisfaction and guilt
   b) an increase in overweight people
   c) less balanced diets than in the past
   d) a population of people who only eat comfort food

After Specs

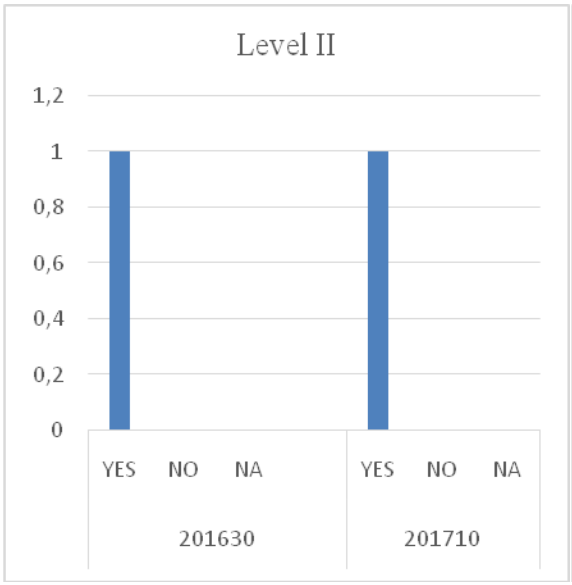


Figure 123

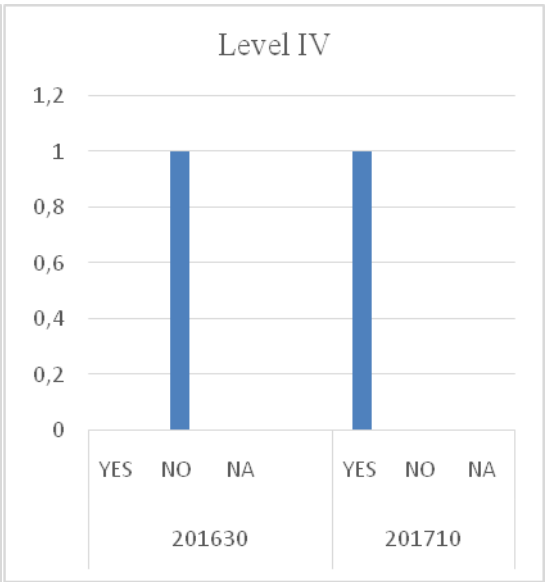


Figure 124

Figure 123 and 124 indicate the text dependency that answers to questions have after Specs. 201630 test of level II and 201710 of both levels II and IV have text dependent answers to the questions asked while 201630 tests of level II does not consider this important aspect due to the fact that questions and answers are not only related to the texts but can be answered by relating to other stems and keys. (See sample)

Sample

DO NOT WRITE ON THIS PAPER---TEST BOOKLET---

Which of these ideas are mentioned in the text? Put a check mark next to the 4 ideas that are discussed in the passage. You will not use all of the sentences. (4 points) .

1. _____ A person who finds inspiration in little things and makes a big business around it.
2. _____ Sometimes, investigating about your possible product is a good idea.
3. _____ Just because someone doesn't use it, does not mean it is outdated or cannot be used again.
4. _____ When all your options are gone, make sure you go to another place to find inspiration.
5. _____ Even if you don't succeed at first, it is always a good idea to try another time.
6. _____ The need to find a better income and a better working schedule are big motivations.

As the example shows, this question is text dependent, since the answers to the questions cannot be deduced from questions, students need to rely on the passage.

5. Are the options of the questions parallel? (formatting)

Before Specs

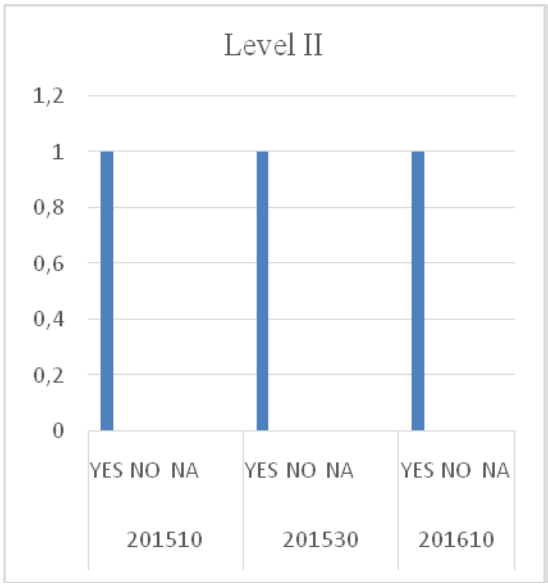


Figure 125

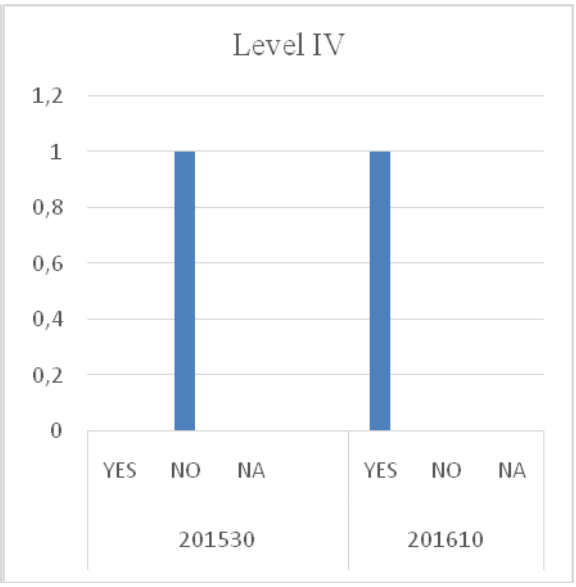


Figure 126

Figure 125 and figure 126 suggest the parallelism of the options of the questions before Specs. *201510 and 201530* tests of level II and *201610* tests of both level II and IV show that the options of the questions are parallel while *201530* test of level IV possess options that are not similar since they do not start with the same part of speech (nouns, adjectives, verbs). Thus, three out of four tests comply with the parallelism of the questions.

After Specs

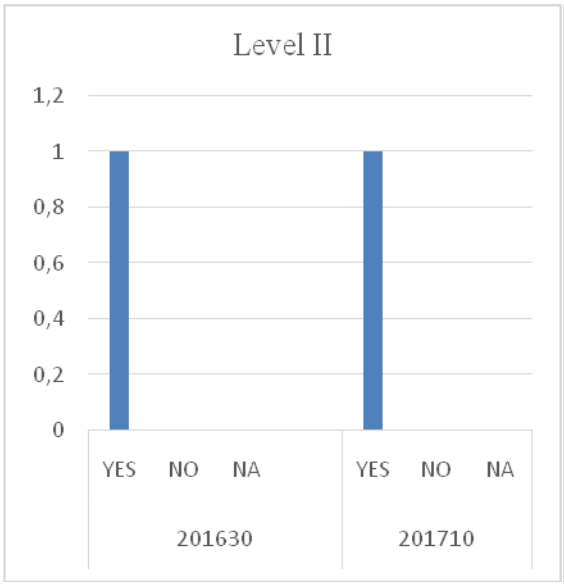


Figure 127

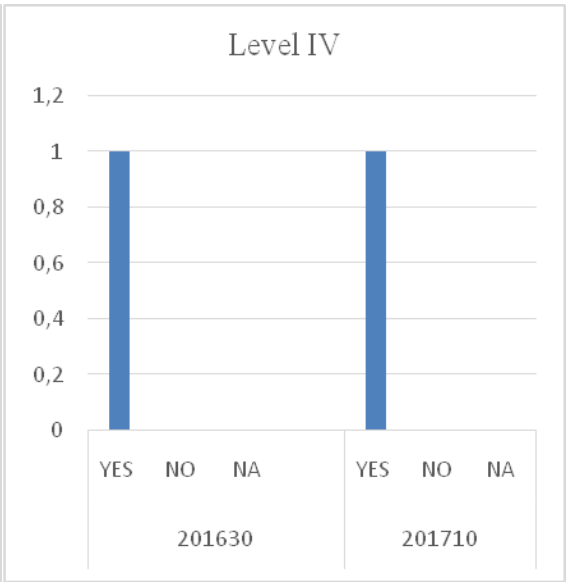


Figure 128

Figure 127 and 128 suggest the parallelism of the options of the questions after Test Specs. *201630 and 201710* tests of both levels II and IV show that the options of the questions are parallel since they start with the same parts of speech (nouns, adjectives, verbs). Thus, all reading tests of levels II and IV comply with the parallelism of the questions. The following example shows how parallel options are.

Part One: Listening One: "Business is a Game"

A. Listen for Main Ideas (QSKL2 CD1 Track 40; 4.28 mins)

Two friends, Moy and Hannah, are talking about an assignment for a business class. The assignment is to play a computer game that teaches some business ideas. You will listen to their conversation one time. Circle the correct answer for each question **on the Answer sheet**. *(4 points; 1 point per question)*

1) What does Moy think about the Lemonade Game?

a) It´s fun, but it can´t help him learn about business.
b) It isn´t very interesting, but it can teach him about business.
c) It´s entertaining and useful for learning about business.

2) Which of these things can you learn from the Lemonade Game?

a) the connection between supply and demand
b) the connection between supply and price
c) the connection between lemons and supply.

3) What happened when Hannah played the game?

a) She made a lot of money.
b) She lost a little money.
c) She made too much lemonade.

4) What is Hannah´s opinion of using a game to learn business?

a) She thinks it is a good way to learn.
b) She thinks it only works for lemonade businesses.
c) She thinks it is not the best idea for a university class.

6. Are the questions formulated through affirmative statements?

Before Specs

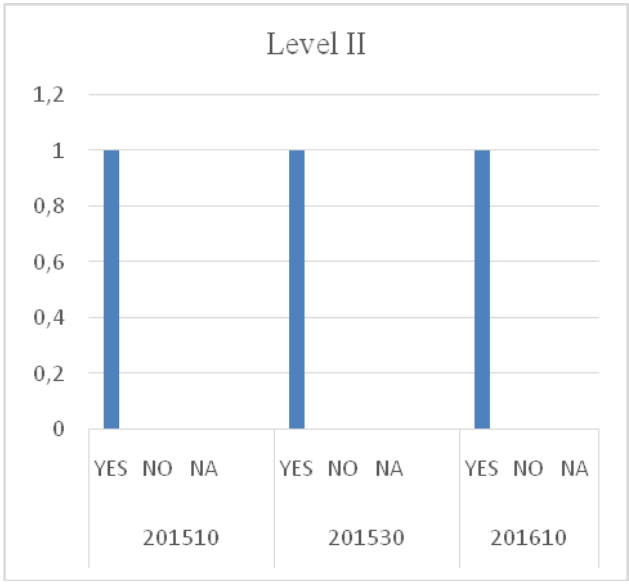


Figure 129

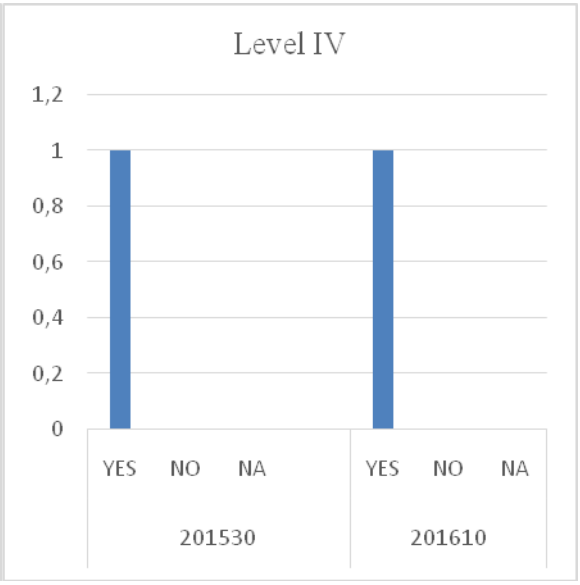


Figure 130

Figure 129 and figure 130 indicate whether the questions are formulated positively or not. This means that having these kinds of questions for students to answer in a test presents a positive impact due to the fact that guidelines instructions were followed. In the *201510*, *201530*, and *201610* tests of level II and *201530* and *201610* tests of level IV the questions are formulated positively. That is to say, every test analyzed contained only items formulated in an affirmative manner.

After Specs

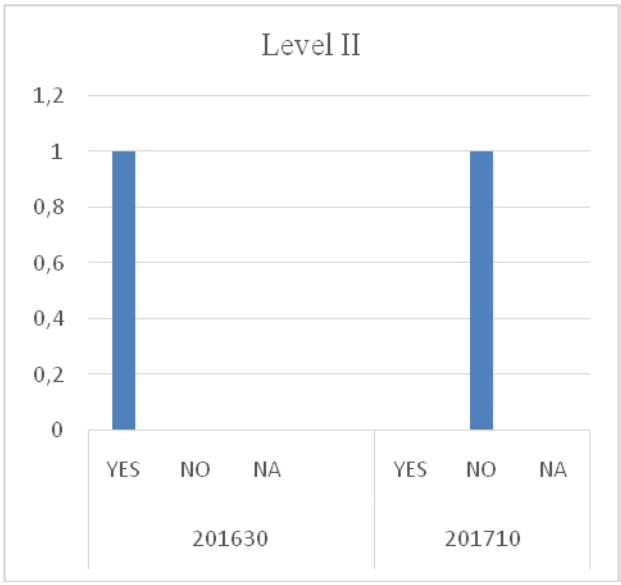


Figure 131

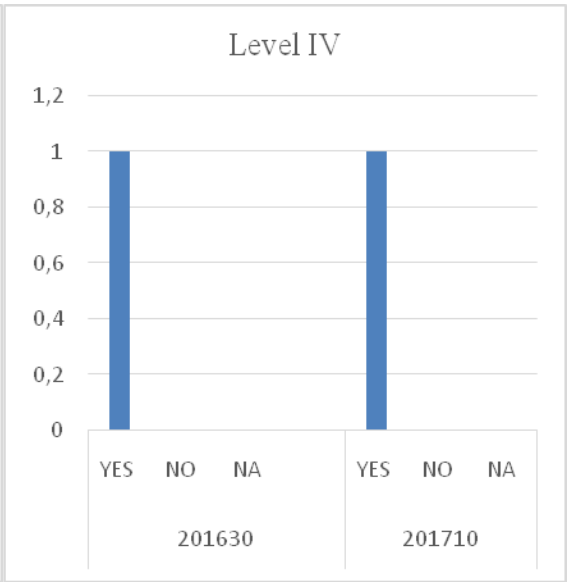


Figure 132

Figure 131 and figure 132 show if the questions are formulated affirmatively. This means that having these kinds of questions for students to answer in a test presents a positive impact due to the fact that Test Specs instructions were followed. In the *201630* test of level II and *201630* and *201710* tests of level IV the questions are designed affirmatively. Unlike, *201710* test of level II has some questions that area asked negatively. That means that three out the four tests analyzed were constructed with items formulated in an affirmative way.

7. Are distractors well designed?

Before Specs

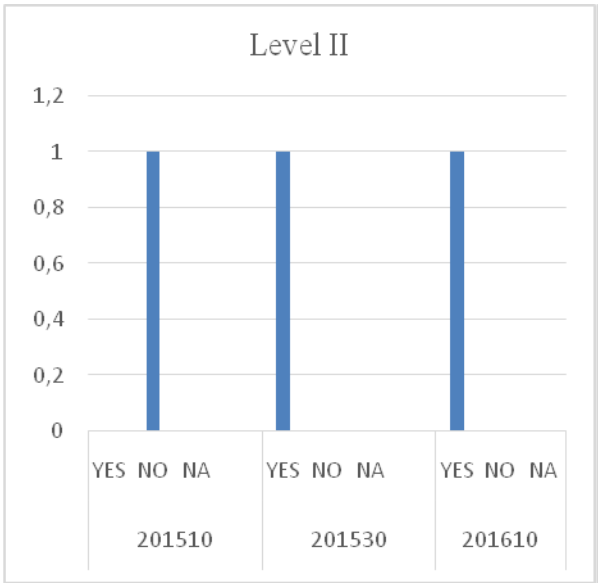


Figure 133

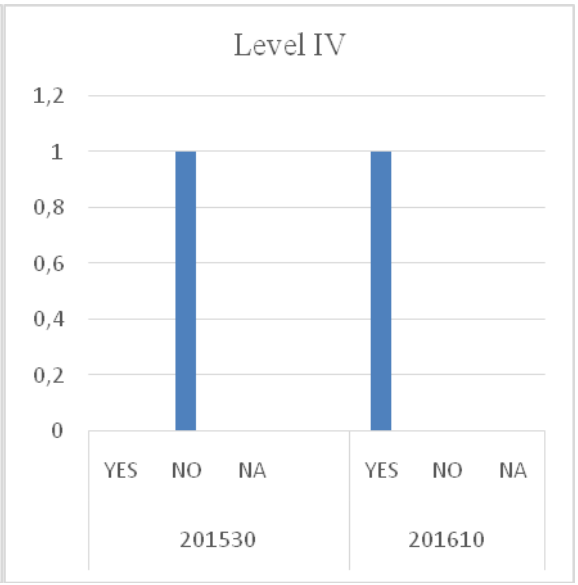


Figure 134

Figures 133 and 134 illustrate the design of distractors in multiple choice questions. Distractors are very important in multiple choice questions because they do not only challenge students to identify what the correct answer is, but they also foster critical analysis in the application of the exams. As presented in the figures, *201530* and *201610* tests of level II and *201610* test of level IV do a good job with distractors, but *201510* test of level II and *201530* test of level IV do not. In other words, the incorrect options in some of the tests analyzed in this study were so obvious that students could easily identify the correct answers without necessarily read the text. (See sample)

Sample

Part Two: Reading Two - On page six, read part two of an article about how culture affects people´s food choices.

A. Read for Main Ideas: Circle the best answer to the following questions.
(4 points; 1 point per question)

1) What is the main idea of Paragraph 4?
   a) Yin and yang help create balance in Chinese cooking.
   b) Yang foods are believed to increase body heat.
   c) Carrots and water are yin foods because they are cold.

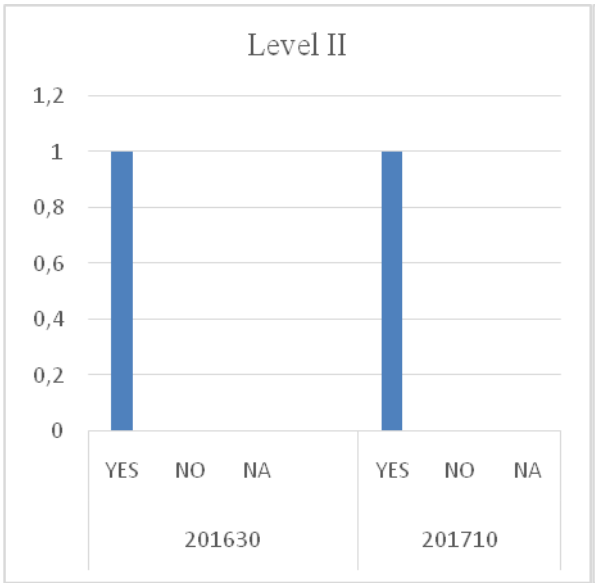
2) What is the main idea of Paragraph 5?
   a) Preparing balanced meals is a challenge.
   b) Too much yang can cause heartburn.
   c) Meals that balance yin and yang can improve health.

3) What is the main idea of Paragraph 6?
   a) China and France have the same concept of a balanced meal.
   b) Culture and food are closely connected.
   c) A balanced meal is about taste and ingredients.

4) What is the main purpose of this complete article **Finding Balance in Food** *(parts one and two)*?
   a) To compare how two cultures find balance in food
   b) To compare why the French do not like fast food and the Chinese do
   c) To compare the concepts of yin and yang to *terroir*

As the example shows, distractors in these questions are not too obvious, that means that students need to rely on the passage to spot the correct answer.

After Specs



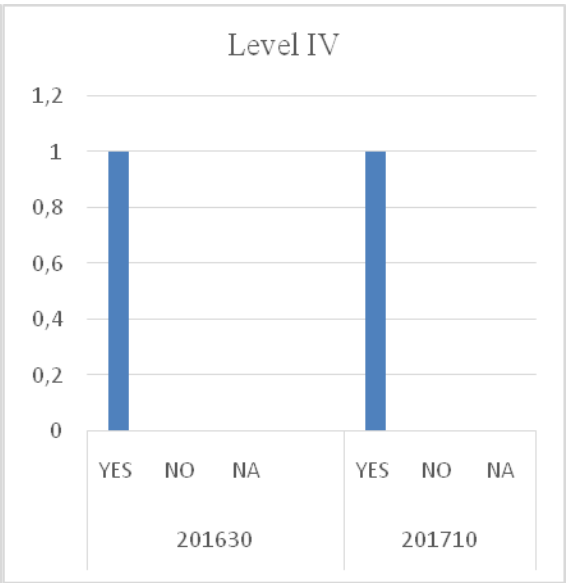


Figure 135 Figure 136

Figures 135 and 136 illustrate whether the design of distractors was accurate in multiple choice questions after Specs. Distractors are very important in multiple choice questions because they do not only challenge students to identify what the correct answer is, but they also foster critical analysis in the application of the exams. As seen in the figures, all tests (*201630* and *201710* of level II and*201630*and *201710* of level IV) do well designed distractors. To be more precise, the incorrect options of the tests examined in this study were not too obvious for test takers, so they actually had to read the text to be able to identify the correct answers. (See sample)

Sample

1) What does Moy think about the Lemonade Game?

a) It´s fun, but it can´t help him learn about business.
b) It isn´t very interesting, but it can teach him about business.
c) It´s entertaining and useful for learning about business.

2) Which of these things can you learn from the Lemonade Game?

a) the connection between supply and demand
b) the connection between supply and price
c) the connection between lemons and supply.

3) What happened when Hannah played the game?

a) She made a lot of money.
b) She lost a little money.
c) She made too much lemonade.

4) What is Hannah´s opinion of using a game to learn business?

a) She thinks it is a good way to learn.
b) She thinks it only works for lemonade businesses.
c) She thinks it is not the best idea for a university class.

8. Are the numbers of questions in chronological order?

Before Specs

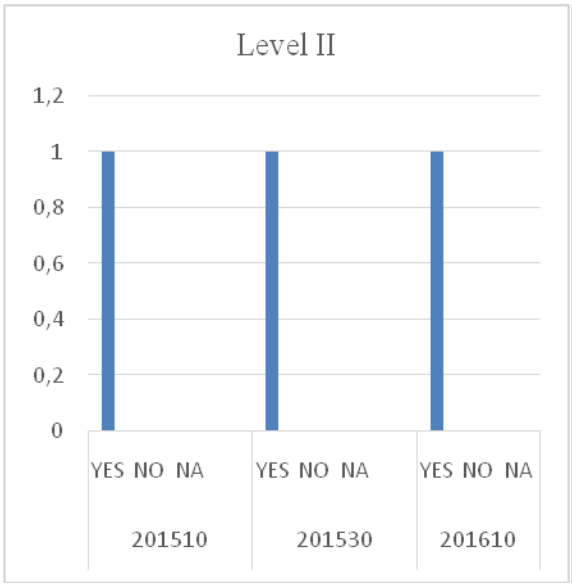


Figure 137

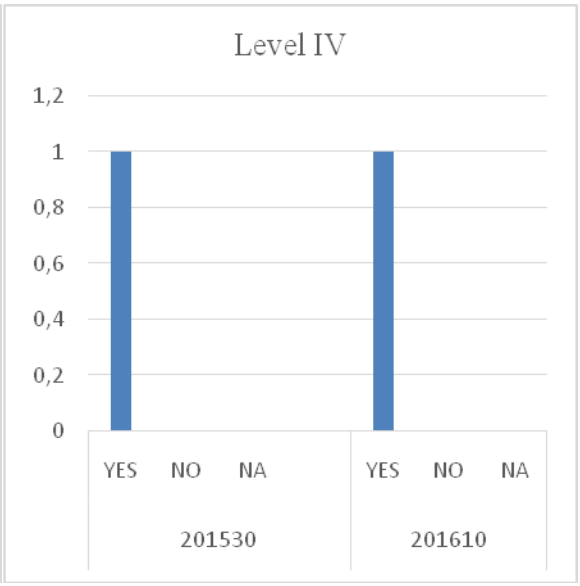


Figure 138

Figures 137 and 138 read if the number of questions is in a correct chronological order. As presented in the figures, the items of the *201510*, *201530*, and *201610* tests of level II and *201530* and *201610* tests of level IV follow a specific chronological order. It means that all the tests analyzed in this study are constructed following a logical order that is fair and not confusing for students.

After Specs

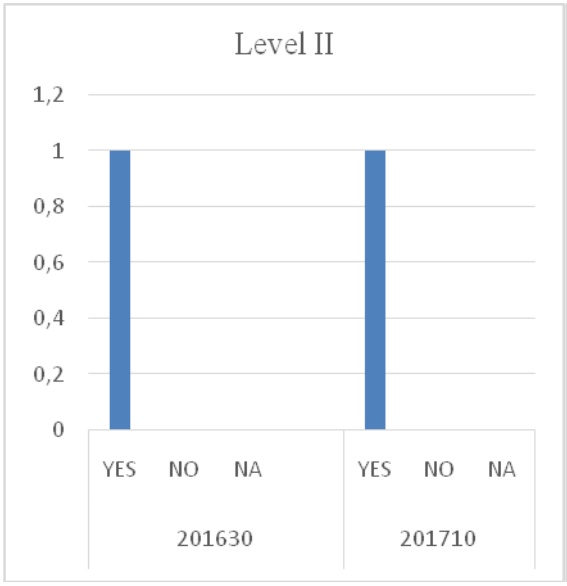


Figure 139

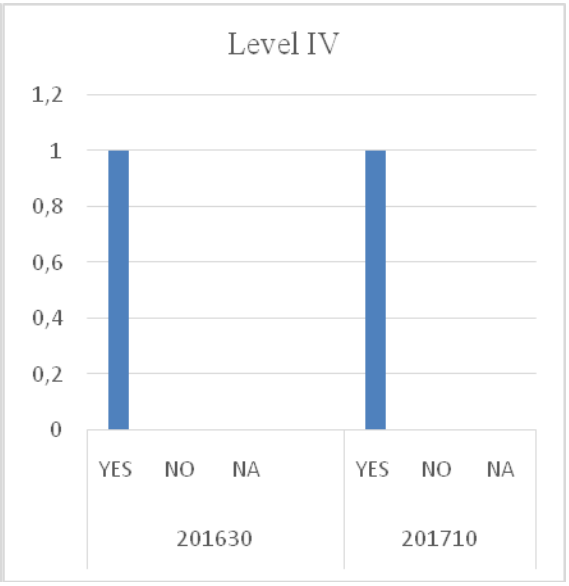


Figure 140

Figures 139 and 140 show whether the number of questions is in a correct chronological order or not. The items of the *201630* and *201710* tests of level II and *201630* and *201710* tests of level IV follow a particular chronological order. This denotes that all the tests examined in this study are designed following a logical order that is clear for test-takers.

9. Do matching exercises have two extra options?

Before Specs

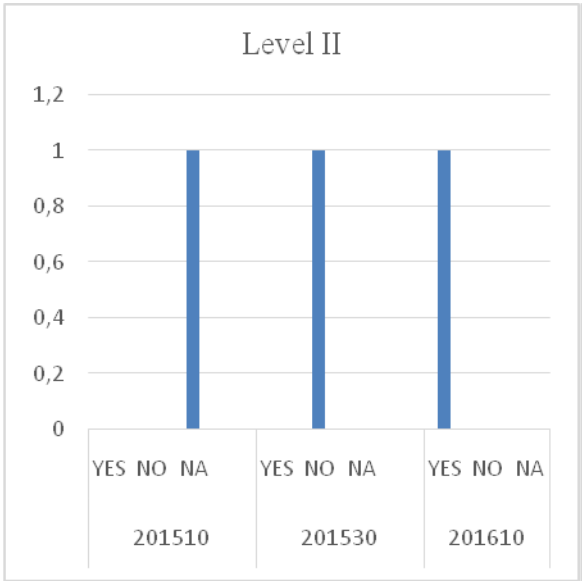


Figure 141

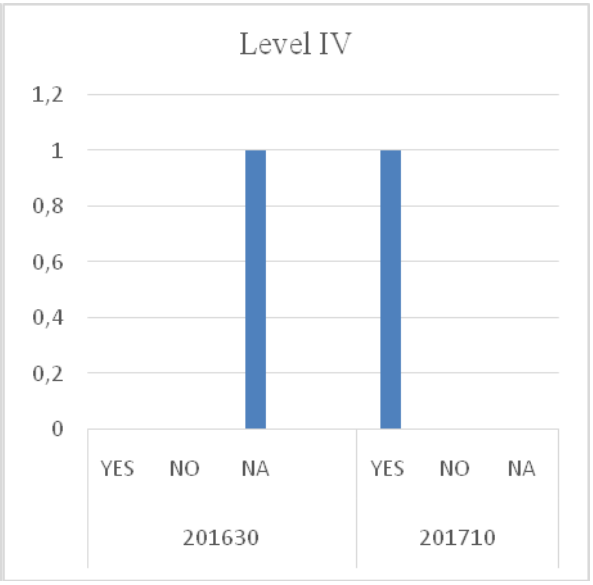


Figure 142

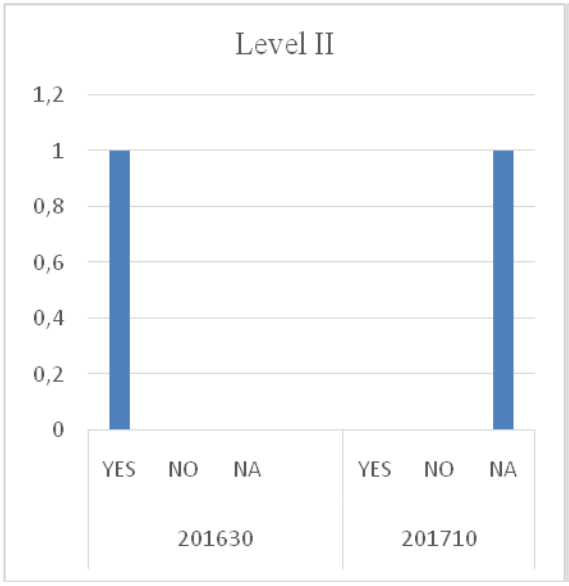
Having extra options in matching exercises is important to prevent students answering items by process of elimination. As noticeable in the figures, *201510* test of level II and *201530* test of level IV do not have any matching exercises. *201610* test of both levels (II and IV) do have two extra choices. On the other hand, the *201530* test of level II does not have any extra options in matching exercises. That is to say, designers of the tests decided not to include matching exercises in two out of the five tests analyzed for this study. (See sample)

Sample

Match the words to their definitions. Write the letter in the box. There are two extra options.
(10 points; 1 point per question)

A	weed	1)	group of plants or animals that rely on each other for food
B	endangered	2)	beginning or origin of something; connection with a place
C	environment	3)	not simple
D	stand up to	4)	established or started in a new location
E	insects	5)	protein in grains, like wheat, that helps make bread rise
F	powerful	6)	habit; practice; duty
G	food chain	7)	protest; defend yourself, family, home, etc.
H	settled	8)	wild plant that people don´t want in their gardens or farms
I	complex	9)	at risk of disappearing
J	roots	10)	natural world; area around us
K	gluten		
L	custom		

After Specs



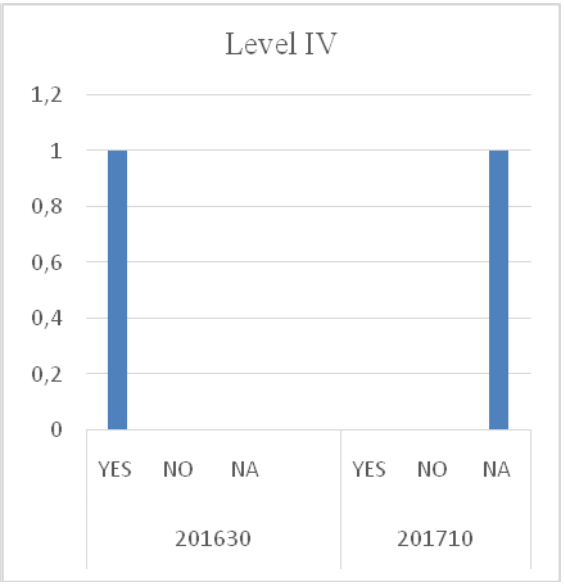


Figure 143 Figure 144

Figures 143 and 144 present information on the matching exercises after Specs. Having extra options in matching exercises is important to prevent students answering items by process

of elimination. As shown in the figures, *201710* test of level II and *201710* test of level IV do not have any matching exercises, whereas *201630* test of both levels (II and IV) do have matching exercises and two extra choices in these exercises. So, teacher-coordinators who are in charge of designing tests chose not to include matching exercises in two out of the four tests studied for this project. (See sample)

Sample

C. INFERRING MEANING FROM CONTEXT
Match each word with its definition. The words in underline bold are from the passage. (8 points)

10. Resource
11. Entrepreneur
12. Retail
13. Concept
14. Stationery
15. Expand
16. Spur
17. Campus

a. Able to make money from new, interesting ideas
b. Products used for writing letters and notes
c. Money, skills, or time that is available when needed
d. To get bigger
e. The land and buildings of a university or college
f. Concerning the sale of things to people in stores
g Motivate, inspire
h An idea about how something is or should be done

The following section will present the analysis made of the tests designed before and after the implementation of the Test Specs.

Tables

The tables presented further on are aimed to present the analysis made of the tests designed before and after the implementation of the Test Specs. They are presented by skill, by level, and by the categories of Validity, Text Language Level Appropriacy, and Test Items Quality. These tables do not only show the amount of times the tests complied or not but they also illustrate the percentages of the times tests met the criteria defined "Yes" or whether they did not "No." These percentages were taken from the relation of the number of questions in each category with the number of times these questions were answered "Yes" and "No."

Listening

Level II Listening

Validity															
After						**Before**									
201710			***201630***			***201610***			***201530***			***201510***			
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	
0	**1**	**3**	**0**	**2**	**2**	**0**	**2**	**2**	**0**	**2**	**2**	**0**	**2**	**2**	
0%	**25%**	**75%**	**0%**	**50%**	**50%**	**0%**	**50%**	**50%**	**0%**	**50%**	**50%**	**0%**	**50%**	**50%**	
Text Language Level Appropriacy															
After						**Before**									
201710			***201630***			***201610***			***201530***			***201510***			
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	
0	**1**	**4**	**0**	**1**	**4**	**0**	**1**	**4**	**0**	**1**	**4**	**0**	**1**	**4**	
0%	**20%**	**80%**	**0%**	**20%**	**80%**	**0%**	**20%**	**80%**	**0%**	**20%**	**80%**	**0%**	**20%**	**80%**	
Test Items Quality															
After						**Before**									
201710			***201630***			***201610***			***201530***			***201510***			
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	
1	**1**	**7**	**0**	**0**	**9**	**0**	**0**	**9**	**0**	**0**	**9**	**0**	**2**	**7**	
11.1%	**11.1%**	**77.7%**	**0%**	**0%**	**100%**	**0%**	**0%**	**100%**	**0%**	**0%**	**100%**	**0%**	**22.2%**	**77.7%**	

Table 1

Table 1 presents the impact before and after the implementation of Test Specs in level II listening tests. The impact is presented in terms of validity, text language level appropriacy, and test items quality. If validity is initially taken into account, 50% of *201510, 201530, 201610* and

201630 tests fulfills with this principle while the other 50% of the tests does not. On the contrary, *201710* listening test of level II shows that only 25% of the tests does not follow the established valid rules, but 75% of those tests does accomplish and fulfill validity guidelines that might make assessment more effective. Concerning text language level appropriacy, before and after Test Specs, 80% of *201510, 201530, 201610, 201630* and *201710* tests reflects the use of adequate listenings in language level whereas 20% of those does not utilize audios that seem to be suitable language level texts. Regarding test items quality, 100% of 201530, 201610 and 201630 comply with the principle of quality because they do construct qualified text items. However, 201510 and 201710 do not completely work on this principle since only 77.7% is formulating worthy questions to be properly assessed.

Level IV Listening

Validity											
After						**Before**					
201710			***201630***			***201610***			***201530***		
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***
0	**0**	**4**	**0**	**0**	**4**	**0**	**1**	**3**	**0**	**0**	**4**
0%	**0%**	**100%**	**0%**	**0%**	**100%**	**0%**	**25%**	**75%**	**0%**	**0%**	**100%**
Text Language Level Appropriacy											
After						**Before**					
201710			***201630***			***201610***			***201530***		
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***
0	**1**	**4**	**0**	**0**	**5**	**0**	**0**	**5**	**0**	**0**	**5**
0%	**20%**	**80%**	**0%**	**0%**	**100%**	**0%**	**0%**	**100%**	**0%**	**0%**	**100%**
Test Items Quality											
After						**Before**					
201710			***201630***			***201610***			***201530***		
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***
1	**0**	**8**	**0**	**4**	**5**	**0**	**1**	**8**	**1**	**3**	**5**
11.1%	**0%**	**88.8%**	**0%**	**44.4%**	**55.5%**	**0%**	**11.1%**	**88.8**	**11.1%**	**33.3%**	**55.5%**

Table 2

Table 2 shows the impact before and after the implementation of Test Specs in level IV listening tests. The impact is presented in terms of validity, text language level appropriacy, and text items quality. Firstly, Listening IV tests have certain similarities and differences when it comes to Validity. For instance, 100% of both 201530 and 201630 tests, before and after Specs,

complies with the principle of validity since tests construct worthy questions and follow the basic validity rules. However, 75% of 201610 properly asks valid questions while the other 25% of them does not. Text language level appropriacy presents surprising results due to the fact that all 201530, 201610, and 201630 tests use appropriate text language level efficiently. Finally, test items quality have different results in the exams. 55.5% of 201530 test does have quality of test items whereas 33.3% of it does not, and 11.1% does not even apply this principle rules. Additionally, the 88.8% of 201610 is ok with the quality of test items, but the 11.1% of this test is not concerning the same principle rules. On the other hand, 201630 test shows that it follows the instructions of test specs in a 55% but the remaining 44% does not follow these guidelines instructions because items are not well qualified.

Reading

Level II Reading

Validity															
After						**Before**									
201710			***201630***			***201610***			***201530***			***201510***			
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	
0	**1**	**3**	**0**	**1**	**3**	**0**	**2**	**2**	**0**	**1**	**3**	**0**	**1**	**3**	
0%	**25%**	**75%**	**0%**	**25%**	**75%**	**0%**	**50%**	**50%**	**0%**	**25%**	**75%**	**0%**	**25%**	**75%**	
Text Language Level Appropriacy															
After						**Before**									
201710			***201630***			***201610***			***201530***			***201510***			
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	
0	**1**	**4**	**0**	**0**	**5**	**0**	**1**	**4**	**0**	**0**	**5**	**0**	**1**	**4**	
0%	**20%**	**80%**	**0%**	**0%**	**100%**	**0%**	**20%**	**80%**	**0%**	**0%**	**100%**	**0%**	**20%**	**80%**	
Test Items Quality															
After						**Before**									
201710			***201630***			***201610***			***201530***			***201510***			
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	
1	**2**	**6**	**0**	**0**	**9**	**0**	**0**	**9**	**0**	**2**	**7**	**1**	**2**	**6**	
11.1%	**22.2%**	**66.6%**	**0%**	**0%**	**100%**	**0%**	**0%**	**100%**	**0%**	**22.2%**	**77.7%**	**11.1%**	**22.2%**	**66.6%**	

Table 3

Table 3 illustrates the influence of Test Specs in level II reading tests. The influence is outlined in terms of validity, text language level appropriacy, and test items quality before and

after Test Specs. From the analysis, when it comes to validity, *201510* and *201530* tests show a tendency to meet the necessary standards to be valid tests. A similar situation is observed in *201630* and *201710* tests. On the contrary, the *201610* test is more distant of being a valid test since only 50% of the questions that evaluate this principle were answered affirmatively.

Concerning text language level appropriacy, in *201510, 201610,* and *201710* tests, 80% of the items that examine the pertinence of texts according to the level, were answered positively. In *201530* and *201630* tests 100% of the questions were affirmatively answered. Regarding test items quality, *201610* and *201630* tests show to have high quality items. Similarly, *201530* test has quality items, 77.7% of questions answered positively affirms that. Contrary, in *201510* and *201710* tests only 66.6% of the questions were affirmatively answered, this shows that the quality of the items in these tests is not high.

Level IV Reading

Validity											
After						**Before**					
201710			***201630***			***201610***			***201530***		
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***
0	**1**	**3**	**0**	**1**	**3**	**0**	**1**	**3**	**0**	**2**	**2**
0%	**25%**	**75%**	**0%**	**25%**	**75%**	**0%**	**25%**	**75%**	**0%**	**50%**	**50%**
Text Language Level Appropriacy											
After						**Before**					
201710			***201630***			***201610***			***201530***		
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***
0	**1**	**4**	**0**	**0**	**5**	**0**	**0**	**5**	**0**	**1**	**4**
0%	**20%**	**80%**	**0%**	**0%**	**100%**	**0%**	**0%**	**100%**	**0%**	**20%**	**80%**
Test Items Quality											
After						**Before**					
201710			***201630***			***201610***			***201530***		
NA	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***	***NA***	***No***	***Yes***
1	**0**	**8**	**0**	**1**	**8**	**0**	**0**	**9**	**1**	**3**	**5**
11.1%	**0%**	**88.8%**	**0%**	**11.1%**	**88.8%**	**0%**	**0%**	**100%**	**11.1%**	**33.3%**	**55.5%**

Table 4

Table 4 reflects the influence of Test Specs in level IV reading tests. The influence is shown in terms of validity, text language level appropriacy and test items quality before and after Test Specs. Evidently, concerning validity, *201610, 201630,* and *201710* tests seem to be valid tests.

In contrast, the *201530* test is more distant of being a valid test since only 50% of the questions that evaluate this principle were answered positively. Regarding text language level appropriacy, in *201530* and *201710* tests, 80% of the questions that analyze if texts are pertinent or not according to the level, were answered affirmatively. In *201610* and *201630* tests, 100% of the questions were affirmatively answered. When it comes to test items quality, *201610* test has high quality items. In a like manner, in *201630* and *201710* tests, items seem to be of high quality. 88.8% of questions answered positively confirms it. Contrary, in *201530* test only 55.5% of the questions were affirmatively answered; this shows that the quality of the items in this test is not high.

7. Discussion and conclusions

Discussion

This chapter presents the discussion of the paper by examining, interpreting, qualifying, and drawing inferences from findings, and it also helps understand the researchers' views. This research aimed to analyze how the use of Test Specs affect test design in terms of the selection of level appropriate texts and the creation of level appropriate questions in the EFL program at private university in Colombia. An Exploratory mixed method was used since the results are analyzed using a qualitative approach and the discoveries are explained and validated using a quantitative one. The analysis is done explaining the main discoveries found before and after the implementation of Test Specs in listening and reading tests of levels II and IV. Taking into account the main theories on which this research is based, the findings are presented next. These findings are divided per level, per skill and before and after Test Specs.

Listening Tests Analysis

Level II- Validity

One of the most important findings is that in none of the listening exams of Level II, before and after Test Specs, the outcome of the level "organize information from multiple texts" (See Appendix A) is not evaluated even though it was stated in the syllabus. This means that the principle of validity here was not completely fulfilled since outcomes stated in the syllabus should be evaluated or measured somehow according to curricular policies in the institution. It was also found that True/False exercises were not allotted the fair amount of points because students had to not only identify if statements were true or false but also correct the false ones, before and after Test Specs. When students have to not only identify whether a statement is true

or false but also they need to write the correction, the latter task requires further analysis, which is why it should be allotted more points upon successful completion. Another positive aspect found in all the tests is that all of them had a reasonable number of questions, this means that students might have the fair amount of questions to be completed in the expected time frame.

Another aspect found after the analysis is that before the use of Test Specs in the design of tests, in two out of the three tests analyzed, the distribution of the items was not suitable and all tests were constructed following the document called Assessment Handbook as a guideline. It is worth mentioning here that some audios and some vocabulary and grammar exercises were recycled from previous tests, which might have an impact on the validity of tests.

Level IV- Validity

Regarding listening tests of level IV before Test Specs, the following information was found: in one out of the two tests examined, it was found that the layout was not coherent and some instructions were not clear. Also the listening strategies such as listening for main ideas, listening for details and make inferences were not explicitly stated in the exam even though they were evaluated. This might interfere with the principle of transparency. It was also observed that there were some explicit grammar and vocabulary exercises in the tests; however, the syllabi did not have explicit grammar and vocabulary outcomes. Some of these grammar and vocabulary exercises in tests were recycled from former tests. Something tests before and after Test Specs had in common is that all they had reasonable number of questions and they were constructed following specific guidelines (Assessment Handbook and Test Specs).

Level II- Text language level appropriacy

When it comes to the appropriacy of the texts for the level, after the analysis of the listening level II test, it was found that all of them had language representative of real world use. They dealt with topics such as *High Tech Health, Doctor's Appointment, Volunteer Vacation*, to name a few. The listening texts were quite long for the level (4'20'', 5'13'', 4'08'', 6'06'') and the number of questions allocated to them was not consistent with the length. For instance, four questions were asked for a four-minute recording. It was also found that in some listening texts, the low speed could disrupt students' attention and also affect the fair and correct evaluation of the outcome "infer speaker's tones, attitudes, and opinions."

Another finding from the analysis made is that before the use of Test Specs one listening text was not related to the topics studied in class, which makes the text problematic because this interferes with the principle of content validity, since the test was not accurately assessing the course content. Also, in one of the tests, in the make inference section, one of the questions was ambiguous to students, it did not seem to have a correct answer; none of the keys was true taking into account the listening text.

Level IV- Text language level appropriacy

The listening tests of level IV before and after Test Specs were designed bearing in mind real world language use, the topics of some listening texts were: *Prenuptial Agreement, Kyoto Protocol, Recycling, and Divorce.* All listening texts were level appropriate, and they had contextualized and unambiguous items and these were also related to the topics of the units studied in class. A specific purpose and language use in mind was visible in the tests and they were also designed for a particular group of test-takers.

Level II- Test items quality

When analyzing quality of the items in each test, the following was found: in one out of the three exams of level II before Test Specs some questions were not text dependent but they could easily be answered with students' prior knowledge and the matching exercise did not have extra options. In contrast, the other two tests did have text dependent items and extra options in matching exercises. After Test Specs, it was observed that two exams had text dependent questions and one of them had extra options in matching exercises and the other one did not have this kind of exercise. In the last test analyzed, it was evident that in the make inference section there were items that were neither clear nor precise.

Some positive aspects common to all tests are: in multiple choice items all keys were clearly identified as the answer to questions asked; the keys were also parallel. This means that they were consistent; they started with the same part of speech. Distractors were also well designed since none of them was too obvious for students to guess the correct answer. All questions were formulated affirmatively and the numbering of questions was correct, to avoid confusing students.

Level IV- Test items quality

The quality of the items was examined in the listening tests of level IV and this was found: before and after Test Specs, students could give the correct answer to some questions not necessarily by listening to the audios but based on their prior knowledge about the topic and also on the information given by other items and keys. Some distractors were too obvious that students could identify the correct answer, again not by listening to the audios, but because of the incorrect ones obviousness. The layout of one of the exams designed after Test Specs might

confuse students, since they had to answer on an answer key and the numbering was not correct. For example, question number one was not on the exam but it was on the answer sheet.

Positive findings common to all tests are: all questions were formulated as affirmative statements and the numbering of questions was correct, to avoid students' confusion. In multiple choice items, all keys were clearly identified as the answer to questions asked, the keys were also parallel, this means that they were consistent, they began with the same part of speech.

Reading Tests Analysis

Level II- Validity

Concerning reading tests of level II before and after Test Specs, it was found that in none of the tests the outcome "differentiate between opinion and fact" was assessed (See Appendix A). This means that the principle of validity may have been affected because the content and outcome were not evaluated in the way they had previously been taught. Another important aspect was that all level II reading tests, before and after Test Specs, had reasonable number of questions that can be completed by students within the expected time frame. Similarly, it is worth mentioning that all level II reading tests were designed following specific guidelines which are the Assessment Handbook (before Test Specs) and Test Specs as detailed instructions to construct valid exams.

Regarding reading tests of level II before Test Specs, it was also found that grammar and vocabulary exercises were evaluated in only one of the three reading tests whereas another test did not distribute items accurately because the reading texts and items were located in different pages of the exam. This can interfere with the face validity since the test's layout is not friendly visible to students. (It may worth mentioning that, as for level IV, there are no explicit grammar

and vocabulary outcomes included in the syllabi; however, these sub-skills are tested). On the other hand, one of the reading tests after Test Specs presented an issue with another outcome. Instead of evaluating the outcome "connect information from multiple texts," it was assessing a complete different outcome, "reading for purpose", even though it was not part of the outcomes of the level. Additionally, all the reading tests, after Test Specs, still used the same text and questions from previous exams which might trigger questions on the validity of the reading assessments.

Level IV- Validity

Furthermore, it was found that Reading tests of level IV, before and after Test Specs, also had certain issues with the outcome "differentiate between opinion and fact." As it can be seen, none of levels II and IV reading tests evaluates this outcome despite the fact that it belonged to the objectives of the language program. Another aspect revealed after the analysis is that one out of the three level IV reading tests, before Test Specs, had no reasonable questions that can be completed by students within the expected time frame. On the contrary, Reading tests of level IV, after Test Specs, presented reasonable number of questions asked in the tests considering a time frame, but it did have drawbacks when assessing certain strategies since main ideas exercises were not properly asked, and detail questions were actually evaluated as main ideas questions.

Positive findings common to all tests are that items were well distributed, and they were all designed following specific guidelines which are the Handbook (before Test Specs) and Test Specs as the detailed instructions to construct valid tests.

Level II- Texts language level appropriacy

When it comes to the appropriacy of texts in reading tests of level II, before and after Test Specs, it was found that all of them have language representative of real world use; the texts dealt with topics such *Finding balance in food, Urban farmer, Riding a bicycle, The climate train*, to name a few. All of these level II tests contained items/tasks that are unambiguous to the test taker, which means they actually asked questions and exercises clear enough for students to understand. On the other hand, reading tests of level II, before and after Tests Specs, had some differences not only because of the guidelines they used (Assessment Handbook, and Test Specs), but also because of other relevant aspects.

To be more precise, one out of the three reading tests of level II, before Test Specs, had a really long reading not suitable for the level. In another exam, the language used in mind did not correspond with the language studied in class; despite the fact that this last mentioned test used real world language use, it did not evaluate the content in the same way it had previously been taught, and this can negatively interfere with the principle of text language appropriacy. It is also worth mentioning that one out of the two reading tests of level II, after Test Specs, did not implement level appropriate readings due to the fact that they dealt with tenses such present perfect in a level II exam, and the wording in the text itself tended to be more complex than the wording presented in the questions.

Level IV- Texts language level appropriacy

Reading tests of level IV, before and after Tests Specs, also had some positive findings in common. All of level IV reading tests were designed bearing in mind real world language use, the topics of some reading texts were *Breaking up, Getting divorce, New jobs, Climate change*

controversy, Tying the knot, Fossil fuel controversy, and College start-ups. All reading texts were level appropriate, and they had contextualized items that were related to the topics of the units studied in class. A specific purpose and language use in mind is visible in the tests, and they were all designed for a particular group of test-takers. On the other hand, one of the reading tests before Test Specs (Assessment Handbook) and one of the reading tests after Test Specs simultaneously presented negative findings. Neither of these two tests, before and after Test Specs, contained items/tasks that were unambiguous to the test taker, which means they did not ask clear questions and exercises for students to easily understand them.

Level II- Test items quality

When analyzing the quality of items in reading tests of level II, many differences were found between the tests that were designed before Test Specs and the tests that were designed after the implementation of Test Specs. One out of the three reading tests of level II, before test Specs, did not have text dependent questions since there was no need to read the text to identify the correct answer, and anyone could easily have answered these questions by their prior knowledge. This same reading test of level II did not use matching exercises as a manner of assessing language proficiency. In a different manner, another reading test of level II, before Test Specs, did have matching exercises to assess language proficiency, but it did not include extra options in the matching exercises. This test also had unclear and imprecise stems that barely indicated what students had to do.

Concerning the quality of items in reading tests of level II after Test Specs, it was found that one of the tests used the same reading and the same questions from a test that was designed before the implementation of Test Specs. This test also formulated questions negatively, and

used really obvious distractors that facilitated students' identification of the right answers. It did not include either matching exercises that assess the language proficiency of the level.

Furthermore, one of the reading tests before Test Specs (Assessment Handbook) and one of the reading tests after Test Specs simultaneously complied with the quality of items since both of them had clear and precise stems, clear keys as the answers to the questions, text dependent questions, parallel keys of the questions, positive questions formulated, well designed distractors, questions in numerical order and matching exercises with two extra options.

Level IV- Test items quality

Reading tests of level IV, before and after Tests Specs, also had some positive findings in common. All of level IV reading tests were designed bearing in mind clear and precise stems that indicated what students had to do, clear keys as the answers to the questions, positive formulated questions and questions were in chronological order. However, if analyzed in detail, it was found that two of the reading tests of level IV, before and after Test Specs, did not have text dependent questions, no distractors, and no parallel questions that started with the same part of speech (nouns, adjectives, verbs).

On the contrary, regardless the use of matching exercises, one of the reading tests before Test Specs (Assessment Handbook) and one of the reading tests after Test Specs simultaneously complied with the quality of items since both of them had clear and precise stems, clear keys as the answers to the questions, text dependent questions, parallel keys of the questions, positive questions formulated, well designed distractors, and questions in numerical order. Just one test of both did not use matching exercises (Last reading test of level IV after Test Specs).

Analysis of tables

Tables in the Results section show the percentages of "Yes", "No", and "NA" (Not Applicable) obtained after analyzing the listening and reading exams of levels II and IV using the checklists. Observing the percentages and the categories analyzed in the checklists, it can be concluded that the Test Specs are not making much difference in the design of tests. With respect to validity, in the listening test of level II, before and after Test Specs the percentages were 50% "Yes" and 50% "No", there was only a small change in the last exam (*201710*), where 75% of the questions was answered positively and 25% of the questions was answered negatively.

With regard to text language level appropriacy, the percentages are as follows: before and after Test Specs, "Yes" was the answer to 80% of questions and "No" was the answer to 20% of the questions. The same situation is presented in the test items quality category, where, before and after Test Specs, 77.7% and 100% of questions was answered "Yes" and 11.1% was answered "No".

In the listening test of level IV, the situation is similar, concerning validity 100% of the questions was answered affirmatively, and this shows a good influence of Test Specs in terms of validity. Language level appropriacy, before Test Specs had 100% of questions answered "Yes" but after Test Specs the percentages change, "Yes" was the answer to 80% of questions and "No" was the answer to 20% of the questions. As regards to items quality, the difference is not quite significant because before Test Specs 88.8% of the answers were "Yes" and 11.1% was "NA." The results after Test Specs are very similar.

In the case of the level II reading exam there is not a consistent impact after the implementation of Test Specs, since the percentages of "Yes" and "No" before and after are almost the same in every test. Regarding the principle of validity (75% Yes, 25% No), text language level appropriacy (80% Yes, 20% No), and test items quality (66.6% Yes, 22.2% No, 11.1% NA).

Likewise, the level IV reading exams analysis reveals that the influence of Test Specs is not very strong on the design of tests. Before Test Specs, respecting validity, the percentages were 75% "Yes" and 25% "No", after Test Specs the result was exactly the same. Regarding the text language level appropriacy, before and after Test Specs, "Yes" was the answer to 80% of questions and 20% of questions was answered "No." In test items quality, it is observable a small change, before Test Specs the percentage of questions answered affirmatively was 55.5% against 33.3% answered negatively and 11.1% answered "NA." After Test Specs, it was 88.8% "Yes" and 11.1% "No."

Conclusions

This chapter provides general conclusions to the paper. Here the main arguments are reviewed, the research question is revised, and implications for teaching are drawn.

Test Specs are the documents that explain in detail how the creation of test tasks and items should be. Specs clarify how to organize the test items, how to do the test layout, how to outline the texts, and what kind of texts to include, and how to make difficult choices in the creation of test materials (Fulcher and Davidson, 2007).

This research intended to see the impact of Test Specs on the design of listening and reading exams of Levels II and IV in the EFL Program in a private university in Colombia.

- Tests Specs are well designed and relevant but when put into practice, they might not be so well executed. Based on the analysis of reading and listening tests of level II and IV, it was observed that Test Specs did not have a consistent impact in the creation of the new exams because designers did not appropriately use them despite the fact that Test Specs instructions were quite clear. For instance, while some reading and listening tests of level II and IV showed positive results with the implementation of Test Specs, others did not have effective results since they did not follow Test Specs instructions accurately. Instead, tests seemed to follow the same criteria used before Test Specs, and some exercises were even recycled from previous tests that followed the Assessment Handbook and not the Test Specs.
- Reading tests of levels II and IV, before and after Test Specs, evaluate the same outcome "differentiate between opinion and fact" (level IV outcome) regardless that the level of difficulty in exams should increase as the levels are higher. This event can negatively

affect the principle of validity since tests are assessing the same course outcome in two different levels. It is worth mentioning that valid exams should have outcomes that suit the levels' needs.

- One of the most important findings is that some listening and reading texts and questions are recycled from previous exams many times. This situation supposes an issue because it interferes with tests' validity, since the same texts could be used but not in consecutive terms and not using the same questions. Texts like "*The Urban Farmer*" are used in four consecutive tests and the same questions are asked. It is important to vary the texts and the questions, not only because this information could be leaked to students after some time but also because questions will be not valid after two or three times using the same texts.

- Another issue found regarding validity is that in the tests of both levels, there are some outcomes that are not evaluated. The missing outcomes are "organize information from multiple listenings" (Level II) and "differentiate between opinion and fact" (Levels II and IV). Tests comply with the principle of content validity when they evaluate the course content and outcomes (Brown, 2004). Having this in mind, further analysis should be given to the possibility of either assessing them in the tests or omitting them as outcomes of the levels.

- Concerning Test Specs, there are some aspects that need further explanation. For example, the approximate number of questions exams should have taking into account that the time frame is fifty minutes. Even though, there is an implicit rule of having fifty points in each exam, the number of questions is not established.

Finally, exploring ways to create stronger formal assessment instruments should be of the interest of educational programs. This paper explored the impact of Test Specs in test design; however, a general conclusion is that in the search for more sound ways of designing tests or means of assessment in general, documents such as Test Specs need to be backed up with extended complementary work since its sole creation does not guarantee significant change. Test Specs creation (and re-creation), implementation, and improvement should be continuously encouraged, followed up, and shared among teachers and teacher-test designers, so in the long run, significant changes in assessment tools design can be observed.

References

Bachman, L. F. (2001). Designing and developing useful language tests. *Experimenting with uncertainty: Essays in honour of Alan Davies*, 109-116.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Ernst Klett Sprachen.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Allyn & Bacon.

Burns, W. (2008). Research Only Matters if You Do Research That matters. *Journal of College Science Teaching*, 37(4), 12.

Butler, S. M., & McMunn, N. D. (2006). *A Teacher's Guide to Classroom Assessment: Understanding and Using Assessment to Improve Student Learning*. Jossey-Bass, An Imprint of Wiley. 10475 Crosspoint Blvd, Indianapolis, IN 46256.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Campbell, J.O., J.R. Bourne, P.J. Mosterman, and A.J. Brodersen. 2002. The effectiveness of learning simulations for electronic laboratories. 91 (1): 81-87.

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education*. Routledge.

Coombe, C. A., Folse, K. S., & Hubley, N. J. (2007). *A practical guide to assessing English language learners*. University of Michigan Press.

Creswell, J.W. 2007. Qualitative inquiry and research design: Choosing among five approaches, 2nd edition. Thousand Oaks, CA: Sage Publications.

Davidson, F., and Lynch, B. K. (2002) *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven, US: Yale University Press, 2001. ProQuest ebrary. Web.

Denzin, N. K., & Lincoln, Y. S. (1994). *Handbook of qualitative research*. Sage publications, inc.

Flores, G. S. (2016). *Assessing English Language Learners: Theory and Practice*. Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London, England & New York, NY: Routledge.

Johnson, B., & Turner, L. A. (2003). Data collection strategies in mixed methods research. *Handbook of mixed methods in social and behavioral research*, 297-319.

Kamil, M. L., Langer, J. A., & Shanahan, T. (1985). *Understanding reading and writing research*. Allyn & Bacon.

López Mendoza, A. A., & Bernal Arandia, R. (2009). Language testing in Colombia: A call for more teacher education and teacher training in language assessment. *Profile Issues in TeachersProfessional Development*, *11*(2), 55-70.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension* (Vol. 1). John Wiley & Sons.

Martin, O. M., Mullis, I. V., & Foy, P. (2015). TIMSS 2015 assessment design. *TIMSS*, 85-99.

National, R. C. S. (2002).Scientific Research in Education. Washington, US: National Academies Press. Retrieved from http://www.ebrary.com

(Shavelson and Towne, 2003, p. 2)

Norman K. Denzin, & Yvonna S. Lincoln. (2005). *The Sage handbook of qualitative research*. Sage.

Seliger, H. W., & Shohamy, E. G. (1989). *Second language research methods*.Oxford University Press.

Shulman, L. S. (1980). *Disciplines of Inquiry in Education, an N Overview*. American Educational Research Association.

Teddlie, C., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. *Handbook of mixed methods in social & behavioral research*, 3-50.

White, E. (2009). *Are you assessment literate? Some fundamental questions regarding effective classroom-based assessment. OnCUE Journal*, *3*(1), 3-25.

Appendix A

Level II and IV Learning Outcomes

Level II learning outcomes by skill:

Reading:

- Comprehend main ideas and details.
- Organize information from multiple texts.
- Identify true details and correct false ones.
- Make inferences to understand ideas not stated directly.

Listening

- Identify main ideas and specific details.
- Organize information from multiple listenings.
- Infer speakers' tones, attitudes, and opinions.

Level IV learning outcomes by skill:

Reading:

- Comprehend main ideas and details.
- Understand and interpret information from facts presented in a text.
- Differentiate between opinion and fact.
- Make Inferences to understand ideas not stated directly.

Listening:

- Identify main ideas and specific details in a conversation or talk.
- Identify reasons for a speaker's opinion in a conversation or talk.
- Recognize markers that signal disagreement and a contrasting opinion.
- Recognize repetition of ideas to emphasize key points of a talk/speech.

Appendix B

Sample of Test Specs

ELP Level II Listening, Vocabulary & Grammar Test Specification

Purpose

- Part of final grade that determines if students will progress to Level III
- Verify students' ability to listen to and understand English-language passages at A2 level
- Verify students' command of vocabulary presented in the textbook
- Verify students' abilities to identify and / or use in context adverbs and expressions of frequency; can and can't; should, ought to, and have to
- Verify that students understand topics covered in textbook listenings

---

Constructs

Listening Constructs

- Listen for main ideas and major points
- Listen for specific details
- Make inferences about speakers' tone, attitudes and opinions based on listening
- Organize information from multiple listenings

Grammar Constructs

- Understand and use in context language patterns to give advice and talk about what is right to do
- Tell accurately how often something is done

Vocabulary Constructs

- Identify the meaning of vocabulary studied
- Use words accurately to express ideas (given the context)

Target Language Use

- Most students are studying English because of central government mandate and undergraduate university requirements for graduation
- Grades for non-credit students not counted in GPA

Test Takers

- Undergraduate students of the university
- From 15 years old to mid-20s, most students between 16 and 20 years old
- Mixed gender
- Variety of major subjects (engineering, psychology, communication and mass media, law students most common in ELP non-credit classes, business administration and accounting students most common in credit classes)
- Main educational goal of most students is to receive an undergraduate degree from the university, most students place more importance on major courses than on language courses
- Vast majority are Colombian (a very small number of students from other countries in Latin America, tiny number from countries outside Latin America)
- Vast majority are native Spanish speakers, most from the Caribbean coast of Colombia
- Different social strata, students from strata 1-3 generally on scholarships
- Many students have very limited exposure to English-speaking cultures, while others have traveled in Latin America and to North America and Europe
- Students generally have access to North American media (film, TV, music), though it is often dubbed into Spanish
- Expected A2 level in English
- Some students are placed in the level based on proficiency test scores
- Some students enter the level after taking previous levels of English (*Nivelatorio*, Level I) and passing courses with a minimum grade equivalent to 60 percent (per university standard)

Usefulness

Reliability

- Inter-rater consistency should be as close to 100 percent as possible
- Questions should be objective, with definitively correct and incorrect responses
- To ensure highest levels of reliability, all instructors will receive verified answer keys prior to grading

Authenticity

- Although texts should not exceed an A2 level of complexity, simplifications should still reflect accurate, real-world English usage

Construct Validity

- Students must achieve at least 60 percent correct responses to items (as determined by university policy) to demonstrate minimal ability

Impact

- Listening: The test should encourage teaching listening strategies, general level-appropriate vocabulary, and close listening in class
- Grammar: The test should encourage classroom focus on form and meaning for newly introduced grammar topics, while encouraging a focus on form, meaning and use for recycled topics
- Vocabulary: The test should encourage students to study vocabulary useful for their overall ability in English, as well as their ability to understand specific vocabulary

Practicality

Design Time

Total test design time for experienced level coordinators familiar with the program should be less than 20 hours, including the following steps:

1. *First draft is created and then submitted to the assessment/program coordinator(s) ***four weeks before test*** administration: less than 10 hours
2. Level coordinator makes changes based on coordinator feedback: less than an hour
3. Corrected first draft is reviewed with instructors ***two weeks before test*** administration: less than two hours (in coordination meetings)
4. Coordinator makes changes based on instructor feedback: less than an hour
5. Coordinator creates a second version of the test based on the first version: less than an hour
6. Coordinator creates answer keys for both versions: less than an hour
7. Final draft and highlighted answer keys are printed, sorted, placed in labeled envelopes and ready for distribution one week before test administration: as much as three hours (based on a level with 750 students and 16 instructors)
8. Coordinator makes adjustments and corrections based on student performance and instructor feedback: less than an hour

Total test review time for program or assessment coordinator ideally should be less than one hour.

Total test review time for instructors should be less than one hour.

Administration Time

- Students have 50 minutes to complete all sections of the test.

Scoring Time

- Scoring should take no more than one hour per 22-student section for an experienced instructor familiar with the program.

Reporting Time

- Departmental reporting (using a standardized spreadsheet with separate section scores) should take no more than 20 minutes per 22-student section

- University reporting (entering aggregate scores into the central reporting system – Aurora) should take no more than 10 minutes per 22-student section

Space

- Design: Secure office space with computers, internet access and printer
- Review: Classroom or laboratory
- Administration: Conducted in the same classroom and at the same time as normal instruction
- Storage: In secure office or locker space

Materials and Equipment

- Students should provide their own pencils, erasers and sharpeners
- Copies may be made in the Idiomas copy center, using the coordinator's copy code

- Copies will be distributed in envelopes provided by secretarial staff in Block I-1 or I-4
- Level coordinators will provide audio files in MP3 format. At this level, no videos will be used for listening assessment.
- Classrooms must have a computer with audio output to speakers.

Personnel

- Level coordinator plans, designs and revises the test
- Program and/or assessment coordinator(s) provide(s) initial feedback
- Instructors provide feedback
- Students take the test
- Instructors administer, grade, provide feedback to students and report results to the department and the university registrar

---

Security*(This section will have further revision)*

No test should be shared with anyone other than the test coordinator and the program coordinator. Teachers should receive only printed copies, let it be, the day before or the same day of the exam.
Answer keys are handed out by the coordinator along with the copy sets in an envelope with each of the teacher's names. The same envelope is to be returned as soon as the teacher has given proper student feedback (no more than 7 days) to the coordinator, or left with the academic secretary.

Test Structure

The test will have three major sections: listening, vocabulary and grammar.

Listening

25 points total (50 percent of the total test score)

Listening One

Listening Passage

- Related to one of the two main topics covered in the listening chapters of the textbook
- 2.5-4 minutes

95 percent or more of vocabulary should match the following criteria:

- Fall within the first 2,000 words of English
- Come from course vocabulary
- Obvious cognates of words students can reasonably be expected to know in Spanish [1]

Use Compleat Lexical Tutor (BNC-COCA framework) to analyze a transcript of the passage.
http://www.lextutor.ca/cgi-bin/vp/comp/

Example Passage

SAMPLE TRANSCRIPT:

Why do some people live longer than others do?
(Length: 2mins 45 secs; Approx. 130 words per minute)

Interviewer: No one knows exactly the reason why some people live longer than others
Why are they so healthy?Is it their diet? Do they go to the gym more than others?
Well, one man is trying to answer these questions
and that man is Explorer and journalist, David McLean
He's currently traveling to places in regions with large numbers of people aged a hundred and over, and asking the question: Why ARE they so healthy? What are they doing that the rest of us aren't?
At the moment, he's walking on the island of Sardinia in Italy. But he's speaking to us right now on the phone.
David, thank you for joining us today.

David: Hello.

Interviewer: So, first of all, tell us why you decided to visit Sardinia.

David: Well, Sardinia
is an interesting place because men live the same amount of time aswomen. That isn't normal for most countries, m en normally die younger.

Interviewer: And does anyone know the reason why people live longer in Sardinia?

[1] Obvious cognates are those that we can easily recognize as having the same meaning in Spanish and English. Example: frequency *frecuencia*Words that are less obvious from their form should not be allowed simply because they are cognate. Example: chief *jefe*Words that are cognate, but which are unlikely to be familiar in Spanish (usually because they occur infrequently, are archaic, or are technical) should also not be allowed.Example: phalarope*falaropa*

David:
There are different ideas about this butpossibly one explanation is that the family is so important here. Every Sunday,
the whole family meets and they eat a huge meal together.Research shows that in countries where people live longer, the family is important, but also on Sardinia the older mother or grandmother often has authority in the family, as men get older,
they have less responsibility in Sardinian culture. So perhaps the older men have less stress, which means they're living longer.

Interviewer:I see, so, do you think people live longer intraditional societies?

David: That's an interesting question. It's true that even on Sardinia the younger generation are eating more food like chips and burgers. Also young people are moving to the city so they aredoing less exercise because of their lifestyle. So it will be interesting to come back to Sardinia in twenty years and see if people are still living longer.

NOTE: ***VoiceBase*** is highly recommended https://www.voicebase.com/- Test designer will need to verify accuracy of transcription since the program may transcribe a few words incorrectly, if they are not clearly pronounced.

Transcripts can also potentially be done/obtained using built-in dictation function on an iPhone or Macintosh in Notes or another text editing app/application. Pause audio occasionally to keep transcription accurate.Consider speech rate and how it may affect students' comprehension when contrasted to their language proficiency level and characteristics of spoken texts they are exposed to in class. As a reference, use the chart below by Tauroza and Allison (1990, p. 39):

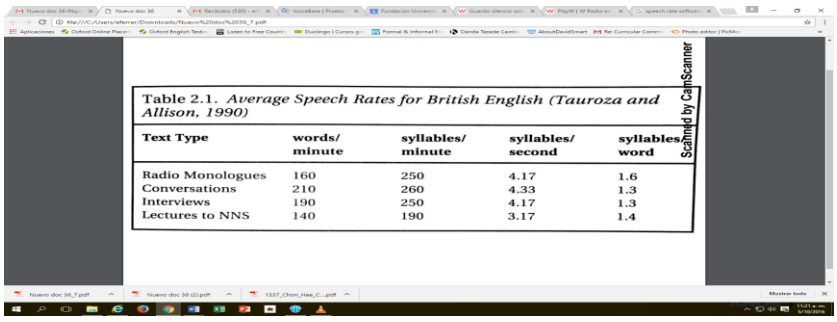


Table 2.1. *Average Speech Rates for British English (Tauroza and Allison, 1990)*

Text Type	words/ minute	syllables/ minute	syllables/ second	syllables word
Radio Monologues	160	250	4.17	1.6
Conversations	210	260	4.33	1.3
Interviews	190	250	4.17	1.3
Lectures to NNS	140	190	3.17	1.4

Scanned by CamScanner

Analysis of this passage (using Compleat Lexical Tutorhttp://www.lextutor.ca/vp/comp/) shows the following:

Currentprofile (token %)		
K-1 (131)	86.75	The following words fall within the K3+ Words: *Explorer, journalist, gym, burger* (these are familiar words to students. They have been covered by the course). Cognates: Italy, authority. The lexical complexity presented by the passage is acceptable.
K-2 (14)	9.27	
K-3 (4)	2.65	
K-4		
K-5 (2)	1.32	
OFF	≈100%	

Syntax

Use **http://www.hemingwayapp.com/**to analyze the text. This tool can help designers identify the syntactic complexity of the text, so necessary adjustments can be made. The grammar expected at this level is that of A2.2. Students who have taken previous courses in the program should have had meaningful exposure to the use of grammar with communicative purposes. Below, there is a list of grammar topics covered per level prior to level 4:

Nivelatorio*: Subject and object pronouns, verb to be, simple present, singular and plural nouns, descriptive adjectives, possessive nouns and adjectives, prepositions of time and place, there is/there are, questions with who, what, where and when, frequency adverbs, present continuous, modals: can/can't.*
Level 1: *Descriptive and possessive adjectives; simple past; should, ought to, and shouldn't for giving advice; comparative adjectives; questions in the simple past; could and would in questions.*
Level 2:*Wh- questions* in *simple present*; *superlatives*; expressing predictions and future plans; *adverbs* and *expressions of frequency*; *can* and *can't*; *should*, *ought to*, and *have to*.

Main Ideas and Main Points

(3-4 points, 1 point per question)

Format must be different from format used for detail questions (i.e. if main idea items are multiple-choice, detail questions may not also be multiple-choice).

Option One: Multiple-choice

- ✓ Stem and options should use familiar vocabulary (2,000-word level or below, obvious cognates and vocabulary from the level).
- ✓ Item order should correspond to the order of the text.
- ✓ Items should be passage dependent (not allow responses based on outside knowledge).
- ✓ Stems will be the question: "What is the main idea of paragraph __?", where the blank corresponds to the number of a paragraph from the passage.
- ✓ Question stems will ask about a main idea (not specific details or inferences) from the passage.
- ✓ Three or four options, with one key.
- ✓ Options will be based on concepts from the passage.
- ✓ Distractors will either describe specific details from the passage, or provide incorrect information. Both types of distractors may be used for one item.
- ✓ Options for each item should be unique. They may not be used for another item in this section.
- ✓ Options should all be of similar length and structure.
- ✓ Students should indicate their response by circling the letter of their intended choice.

Example:

> **A. Listen for Main Ideas and listen for Main Points:** Circle the best answer to the following questions *(2 points; 2 point per question).*

1. What is the main idea of the passage?
   a) Each of the 22 regions in France has its own *terroir*.
   b) In France, balance comes from eating foods from different regions.
   c) The French find pleasure in eating at different kinds of restaurants.

Comments

- ❏ The distractors for item one (options a and c) are specific details.
- ❏ The distractors for item two provide incorrect information (option b) and a specific detail (option c).

2. What is the main point made by the interviewee?
   a) People who have a big family may live longer.
   b) People who live in islands normally eat healthily.
   c) People who live in traditional societies die young.

Option Two: Check correct options

- Stem and options should use familiar vocabulary (2,000-word level or below, obvious cognates and vocabulary from the level).
- Options should **not** correspond to the order of the text.
- Options should be passage dependent (not allow responses based on outside knowledge).
- The stem will be a statement about possible main ideas: "Three main ideas from the passage are:"
- Five options, with three keys
- Distractors will either describe specific details from the reading passage, or provide incorrect information. Both types of distractors may be used.
- Options should all be of similar length and structure
- Students should indicate their response by writing a check mark in a blank to the left of their intended choice.

Example:

A. Read for Main Ideas: Put a check (√) next to the three main ideas from the passage. *(3 points; 1 point per item)*

1. Three main ideas from the reading are:
   ___ In France, balance comes from eating foods from different regions.
   ___ The French find pleasure in eating at different kinds of restaurants.
   ___ The French balance their meals by serving many small courses.
   ___ French people prefer simple meals and they like to eat quickly.

___ Culture has an effect on people's ideas about a balanced diet.

Comments

- ❏ The second option is a distractor based on specific details.
- ❏ The fourth option is a distractor based on incorrect information.

Option Three: Matching

- Options should use familiar vocabulary (2,000-word level or below, obvious cognates and vocabulary from the level).
- Options should **not** correspond to the order of the text.
- Options should be passage dependent (not allow responses based on outside knowledge).
- There should be one more option than speakers or passages: four options with three keys or five options with four keys.
- Distractors will either describe specific details from the passage, or provide incorrect or unrelated information.
- Options should all be of similar length and structure.
- Students should indicate their response by writing a number (1-3 or 1-4) in a blank to the right of their intended choice.

Example:

A. Listen for Main Ideas: Listen to speakers 1 (S1), speaker 2 (S2), and speaker 3 (S3). Then, match the main points below 1 through 4 with the right speaker. **You will <u>not</u> use one option.***(3 points;1 point per question)*

1. ___ In France, balance comes from eating foods from different regions. a) S1
2. ___ The French balance their meals by serving many small courses. b) S2
3. ___ Culture has an effect on people's ideas about a balanced diet. c) S3
4. ___ The French find pleasure in eating at different kinds of restaurants.

Comments

- ❏ The fourth option is a distractor based on specific details.

Details

(4-5 points, 1 point per question)

Option One: Multiple-choice

- Stems may be a question or an incomplete sentence.
- Question stems will ask about a specific detail (not main ideas or inferences) from the passage.
- Incomplete sentence stems will state a detail (not main ideas or inferences) from the passage. The portion to be completed will be about a specific detail.
- Three or four options, with one key.
- Options should all be of similar length and structure.
- Stem and options should use familiar vocabulary (2,000-word level or below, obvious cognates and vocabulary from the level).
- Item order should correspond to the order of the text.
- Items should be passage dependent (not allow responses based on outside knowledge).

Option Two: True/False, make false statements true

- Each stem will be a single, short statement that is clearly true or false.
- Items should be passage dependent (not allow responses based on outside knowledge)
- Items should use familiar vocabulary (2,000-word level or below, obvious cognates and vocabulary from the level)
- Item order should correspond to the order of the text
- False items should be created by taking a statement that is true about the reading passage, and changing a word or phrase to make the statement false:
    - Opposites
    - Different information or meaning
- False items should not allow correction by inserting or removing the words no or not — with or without an auxiliary — to change the meaning. Questions that allow this kind of response may not demonstrate that the student has genuinely

understood the reading.

- False information in the statement has to be underlined by students. ONLY the words that make the statement incorrect can be indicated.

- The instructions should clearly state that students are to circle either True or False, and the words True and False should appear to the left of each item. This is to make answering faster for students and to prevent students using Spanish words or abbreviations to answer.

Example:

Listen for Details: Read the statements below. Then, listen to the passage and decide if theyare true or false. Circle True or False. If the statement is false, underline the false information given in the statement.
(4 points; 1 point per question)

Example:

True False *Terroir* means international food traditions.
Terroir means local food traditions.

1. True False There are no fast-food restaurants in France.
There are fast-food restaurants in France. OR Many fast food restaurants have changed their menu.

2. True False A person´s culture has an effect when they are trying to find balance with food.
_____

3. True False The French eat small portions of food.
_____

4. True False French meals always have seven courses.
French meals can have three to seven courses/sometimes have 7 courses.

Option Three: Short answer

- Each stem will be either a single, short statement with a piece of information missing, or a question which can be answered by giving a detail from the passage.
- Items should be passage-dependent (not allow responses based on outside knowledge).
- Items should use familiar vocabulary (2,000-word level or below, obvious cognates and vocabulary from the level).
- Item order should correspond to the order of the text.
- Correct responses should be from one to four words long.

Example:

Listen for Details: Read each question below. Then, write a key word that answers each question.
(4 points; 1 point per question)

Question?	Answer – Key word(s)
Where did the accident happen?	
When did the accident happen?	
Why did the accident happen?	
Who was the victim?	

Inferences

2-4 points, 2 points per question

Option One: Agree or Disagree

- Stem and options should use familiar vocabulary (2,000-word level or below, obvious cognates and vocabulary from the level).
- Options should **not** correspond to the order of the text.
- Options should be passage dependent (not allow responses based on outside knowledge).
- Two options, with one key
- The stem will describe the attitude, opinion, or feelings of people from the

passage. The options will be either disagree or agree.

- Students should indicate their response by circling agree or disagree.

Example:

C. Make Inferences: Read the comments. Then listen to the passage and indicate whether you think these people probably agree or disagree with the statement? Circle agree or disagree. *(4 points;2 points per question)*

1. French people:"We like the simplicity of traditional fast food." disagree agree

2. French people: "Enjoying the taste of your food is very important." disagree agree

Option Two: Check the correct option (multiple-choice)

- Stem and options should use familiar vocabulary (2,000-word level or below, obvious cognates and vocabulary from the level).
- Options should be passage dependent (not allow responses based on outside knowledge)
- The stem will be of the form "Which fact can you infer from the passage?"
- Three options, with one key
- Distractors will state facts which cannot be inferred from the passage or are incorrect inferences.
- Options should all be of similar length and structure
- Students should indicate their response by writing a check mark in a blank to the left of their intended choice.

Example:

C. Make Inferences: Read the following statements. Then, listen to the passage and check (√) the inference that can be made from each statement. *(4 points; 2 points per question)*

1) French people

__√__ Enjoying the taste of your food is very important.

____ We like the simplicity of traditional fast food.

____

2) Nutritionists

__√__

____ There is only one way to eat a balanced diet.

____

Listening Two

Listening Two:this listening passage ***should not be connected*** to the listening passage in the first section.
For Listening Two, designer may use any of item design options proposed for ListeningOne that were not chosen for the exam up to this point.

General comments:

More than two passages can be used.
If multiple passages are used, designer needs to follow the recommendations indicated in this document for text selection and item design and needs to verify that the reproduction of passages and students time to answer fits time allotted for the exam.

Vocabulary

The vocabulary section includes words from book lists and general-at-level vocabulary inferred from context.
(This section is 25% of the total test score, 1 point each)

Matching

(10 items, 1 point each)

- Answers are always marked correct or incorrect, no partial points
- If a student matches an option to multiple items, each item must be marked incorrect. Asking the student which answer was intended at a later time compromises the reliability of the test.

- All items and options should be in English.
- There will be ***two or three additional options*** (distractors) to prevent answering items by process of elimination.
- If a key can be recognized by its form, then there must be a plausible distractor of the same form (examples: proper nouns, numbers, plural nouns, adverbs ending in *-ly*, etc.)
- Options consist of one word or phrase, taken from target vocabulary in the textbook
- Stems are definitions taken directly from the textbook
- Stems and options should be distributed randomly. No more than one option should appear directly across from the matching stem.
- The key for each item must match the part(s) of speech, meaning(s) and use(s) as presented in class, based on the textbook.

  Example 1: The word *protest* is presented in NorthStar 2 as a verb, and is used as a verb in the texts. Therefore, in the vocabulary section of the test, the word *protest* should match a definition for the verb, not the noun.

  Example 2: The word *roots* is presented in NorthStar 2 in a cultural context, as a synonym of *heritage* or *origin*. Therefore, in the vocabulary section of the test, the word *roots* should match a definition with this meaning, and <u>not</u> one meaning the roots of a plant, the roots of hair, etc.
- Different versions of the test for one administration should have the same items, but both the stems and the options should be in a different order.

Formatting should match the following example:

Instructions:Match the definitions on the left with the words that they define on the right.

	a. stairs
1. ___ the day of the week after Sunday and before Tuesday	b. realice
2. ___ having a strong desire to know about something	c. divide
3. ___ a set of steps between two floors of a building	d. Wednesday
4. ___ to be aware of sounds with your ears	e. foreign
5. ___ a largeamount	f. Monday

6. ___ to become aware of a fact or situation
7. ___ having a lot of money
8. ___ make something separate into parts
9. ___ in or from a country that is not your own
10. ___ to cook in an oven without fat or liquid

g. curious
h. plenty
i. bake
j. ramps
k. wealthy
l. hear

Option Two: Sentence Completion or Gap Fill Items

- Answers are always marked correct or incorrect, no partial points
- If a student completes different gaps with the same word or phrase, each item must be marked incorrect. Asking the students which answer was intended at a later time compromises the reliability of the test.
- There will be a word bank with the possible answers with two or three additional options (distractors) to prevent completing the gaps by process of elimination.
- Avoid using absurd distractors as they do not contribute to the test.
- If a key can be recognized by its form, then there must be a plausible distractor of the same form (examples: proper nouns, numbers, plural nouns, adverbs ending in -ly, etc.).
- Options consist of one word or phrase, taken from target vocabulary in the textbook.
- The expected response should be clear from the sentence. Provide sufficient context in the stem.
- Avoid providing grammatical clues in the sentence.
- Formatting should match the following example:

Instructions: fill in the blanks with the most suitable word from the box. Words are not repeated.
Sentence level

allergic	resist	approve	environment	insects

1. In our restaurant, the chef controls everything. She needs to _____ every plate that goes to the customer.

2. If Betsy eats peanuts, she has trouble breathing and needs to go the hospital immediately. She is _____ to peanuts.

3. Something is eating the tomatoes in my garden! I am not sure if it is birds or _____.

(From: Coombe, 2011)

Grammar

The grammar section will have three sub-sections, covering the one topic recycled from the previous level and two grammar topics introduced in this level.

(This section is 25% of the total test score).

• Items will require that demonstrate correct usage of Standard English structures.
• Items will require students to identify and / or use in context adverbs and expressions of frequency; can and can't; should, ought to, and have to.
• Items in this section should be balanced between identification and production. Special attention is to be paid in usage of verb forms, subject/verb agreement and accuracy of response.
Designers can choose from the options described in this document to create items for the grammar section following the established criteria:

- ➢ Matching
- ➢ Gap filling
- ➢ Sentence construction
- ➢ Multiple choice
- ➢ Error correction

Appendix C

Checklists

Level Two Listening assessment

Exam version A_ **Reading exam**

Exam Date April 21st 2015 **Listening exam X**

I. **VALIDITY**	**YES**	**NO**	**NA**	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		It does not have "Organize information from multiple listenings"
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			37 questions
3. Are the items well distributed to make the test valid?		X		True and false exercises should have more marks.
4. Does test construction follow specific guidelines?	X			Assessmenthandbook.
II. **TEXTS LANGUAGE LEVEL APPROPRIACY**	**YES**	**NO**	**NA**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			Volunteers vacations (Machu Pichu), assignment for a business class, a report on why people choose to live where they do.
2. Are the items level appropriate?		X		Listening #2 and # 3 are too long for the level (4:28- 4:51). The low speed might interrupt students' attention.
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			Part 3 seems not to be related to the syllabus. (topics of the level)
III. **TEST ITEMS QUALITY**	**YES**	**NO**	**NA**	**COMMENTS**
1. Is the stem clear and precise (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)		X		Listening # 3 has a question that can be

				answered with information presented in another question.
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?		X		

Level Two Listening Assessment

Exam version A | **Reading exam**

Exam Date Oct 5th2015 **Listening exam X**

I. **VALIDITY** NA	YES	NO		COMMENTS
1. Does the test accurately measure what it intends to measure?		X		It does not have "Organize information from multiple listenings" This test includes grammar and vocabulary.
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			44 questions
3. Are the items well distributed to make the test valid?		X		True and false exercises should have more marks. The number of questions per section is not even.
4. Does test construction follow specific guidelines?	X			Assessmenthandbook.
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA	YES	NO		COMMENTS
1. Is the language in the test representative of real-world language use?	X			How games relate to real life, descriptions of cities.
2. Are the items level appropriate?		X		Listening #2 is too long for the level (6:06). The low speed might interrupt students' attention.
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			Part 3 seems not to be related to the syllabus.
III. **TEST ITEMS QUALITY** NA	YES	NO		COMMENTS
1. Is the stem clear and precise(It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			

4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?	X			

Level Two Listening Assessment

Exam version A | **Reading exam**

Exam Date April 27th 2016 **Listening exam X**

I. **VALIDITY** NA		**YES**	**NO**	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		It does not have "Organize information from multiple listenings" This test includes grammar and vocabulary.
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			49 questions.
3. Are the items well distributed to make the test valid?		X		True and false exercises should have more marks considering that students have to correct the false statement. The number of questions per section is not even.
4. Does test construction follow specific guidelines?	X			Assessment handbook.
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA		**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			How games relate to real life, food tasters talking about their jobs.
2. Are the items level appropriate?		X		Listening #2 is too long for the level (4:47). The low speed might interrupt students' attention.
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY** NA		**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the	X			

question asked?				
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?	X			

Level Two Listening assessment
Exam version A
Exam Date OCT 18TH 2016

Reading exam
Listening exam X

I. **VALIDITY**	**YES**	**NA**	**NO**	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		It does not have "Organize information from multiple listenings"
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			48 questions
3. Are the items well distributed to make the test valid?		X		True and false exercises should have more marks.
4. Does test construction follow specific guidelines?	X			Test specification guidelines
II. **TEXTS LANGUAGE LEVEL APPROPRIACY**	**YES**	**NA**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			Assignment for a business class, doctor's appointment.
2. Are the items level appropriate?		X		Listening #1 is too long for the level (4:28). The low speed might interrupt students' attention.
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY**	**YES**	**NA**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?	X			

Observations:

- The exercises and the audio in the first listening were previously used. (before test specifications: 2015-10, April 21st 2015)
- The vocabulary exercise was previously used in another exam (before test specifications, April 27 2016)

Level Two Listening assessment
Exam version A **Reading exam**
Exam Date March 2017 **Listening exam X**

I. **VALIDITY**	**NA**	**YES**	**NO**	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		The test does not seem to evaluate the outcome "organize information from multiple listenings" It is difficult to infer from speaker's tones.
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			30 questions
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			TESTS SPECS
II. **TEXTS LANGUAGE LEVEL APPROPRIACY**	**NA**	**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			doctor's opinion, high-tech health
2. Are the items level appropriate?	X			
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?		X		Make inferences questions are ambiguous because question #1 does not have a correct answer.
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY**	**NA**	**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)		X		Make inferences (listening #1, question #2)
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			

9. Do matching exercises have two extra options?			X	It does not have matching exercises

Observations:

- It might be necessary to modify listening #1 speed
- Make inferences questions are ambiguous because questions #1 does not have a correct answer and students can answer questions #2 because of the information they listen and not because of the tone of voice since her tone of voice is always plain.
- The test does not seem to evaluate the outcome "organize information from multiple listenings"

Level Two Reading Assessment
Exam version A | **Reading exam X**
Exam Date FEB 24TH 2015 **Listening exam**

I. **VALIDITY** NA		YES	NO	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		It doesn't have "Differentiate between opinion and fact"
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			31 questions
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			Assessment handbook
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA		**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			Finding balance in food, riding a bicycle, the climate train
2. Are the readings level appropriate?		X		Reading #1 is too long for the level although it is divided.
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY** NA		**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)		X		Reading #3 (main ideas): options are obvious.
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?		X		Reading #3 (main ideas): detractors are obvious.
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?			X	

Level Two Reading Assessment
Exam version A | **Reading exam X**
Exam Date AUGUST 20TH 2015 **Listening exam**

I. **VALIDITY** NA		YES	NO	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		It doesn't have "Differentiate between opinion and fact."
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			47 questions
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			Assessment handbook
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA		**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			Food, the urban farmer,
2. Are the readings level appropriate?	X			
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose, a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY** NA		**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)		X		Reading #1 (details) questions 3& 4 don't clearly indicate what students have to do.
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?		X		It doesn't have two extra words.

Observations:
This test has grammar and vocabulary exercises

Level Two Reading Assessment
Exam version A | **Reading exam X**

Exam Date March 10th2016 _____ Listening exam

I. **VALIDITY**	**NA**	**YES**	**NO**	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		It doesn't have "Differentiate between opinion and fact."
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?		X		44 questions
3. Are the items well distributed to make the test valid?		X		Readings and items are not together.
4. Does test construction follow specific guidelines?		X		Assessmenthandbook
II. **TEXTS LANGUAGE LEVEL APPROPRIACY**	**NA**	**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?		X		Food
2. Are the readings level appropriate?		X		
3. Does the test have items that are contextualized rather than isolated?		X		
4. Does the test contain items/tasks that are unambiguous to the test-taker?		X		
5. Is the test developed with a specific purpose and a particular group of test-takers?			X	The language use does not correspond with the language studied in class.
III. **TEST ITEMS QUALITY**	**NA**	**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)		X		
2. Is each key clearly identified as the answer to the question asked?		X		
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)		X		
4. Is the answer to each question text dependent (it does not depend on other stems and keys)		X		
5. Are the keys of the questions parallel? (formatting)		X		
6. Are the questions formulated positively?		X		
7. Are distractors well designed?		X		
8. Are the numbers of questions in order?		X		
9. Do matching exercises have two extra options?		X		

Observations:
Questions and the text are not close enough (page 2, page 6)

Level Two Reading Assessment
Exam version A | **Reading exam X**
Exam Date Sept 9th 2016 **Listening exam**

I. **VALIDITY**	**YES**	**NO**	**NA**	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		It doesn't have "Differentiate between opinion and fact."
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			44 questions
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			Test specification
II. TEXTS LANGUAGE LEVEL APPROPRIACY	**YES**	**NO**	**NA**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			Xavante people Urban farmer
2. Are the readings level appropriate?	X			
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. TEST ITEMS QUALITY	**YES**	**NO**	**NA**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?	X			

Level Two Reading Assessment
Exam version B | **Reading exam X**
Exam Date 201710 **Listening exam**

I. **VALIDITY**	**YES**	**NO**	**NA**	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		Reading #1 question 4 seems to evaluate "making inferences" rather than main ideas
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			30 questions
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			TEST SPECS
II. **TEXTS LANGUAGE LEVEL APPROPRIACY**	**YES**	**NO**	**NA**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			Urban Farmer
2. Are the readings level appropriate?		**X**		
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY**	**YES**	**NO**	**NA**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?		X		Only two questions are formulated negatively.
7. Are distractors well designed?		X		Vocabulary section has very obvious distractors
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?			X	It does not have matching exercises

Observations:

- The test does not seem to evaluate the outcome "Organize information from multiple texts".
- The wording in the text may be more complex than the wording in the questions.

(Question #4, reading 1)

- The text uses present perfect structure in level 2.
- Reading 2 “the urban farmer” is still the same text from those exams that were designed using the handbook.
- Most of the questions from reading #2 are the same (Making inferences and reading for main ideas and details)
- Reading for purpose is not an outcome

Level Four Listening Assessment
Exam version A | **Reading exam**
Exam Date 201530 **Listening exam _X_**

I. **VALIDITY** NA	YES	NO	COMMENTS
1. Does the test accurately measure what it intends to measure?	X		
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X		Listening #1 & #2 are quite long. (4:20 min – 5:13 min). The first listening exercises require more time and have more questions than the last listening exercises.
3. Are the items well distributed to make the test valid?	X		42 questions
4. Does test construction follow specific guidelines?	X		Assessment handbook
II. TEXTS LANGUAGE LEVEL APPROPRIACY NA	**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X		Responsibility, the great banana race, the envelope, prenuptial agreement.
2. Are the listenings level appropriate?	X		
3. Does the test have items that are contextualized rather than isolated?	X		
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X		
5. Is the test developed with a specific purpose and a particular group of test-takers?	X		
III. TEST ITEMS QUALITY NA	**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X		
2. Is each key clearly identified as the answer to the question asked?	X		Main ideas, question #1: The answer is obvious after listening to the exercise
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)		X	Listening #4 can be answered by prior knowledge
4. Is the answer to each question text dependent (it does not depend on other stems and keys)		X	In part #3, some of the keys might give the answer to some of the questions.

5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?		X		Main ideas, question #1: The answer is obvious after listening to the exercise
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?			X	

Level Four Listening Assessment
Exam version | **Reading exam**
Exam Date 201610 **Listening exam X**

I. **VALIDITY** NA		YES	NO	**COMMENTS**
1. Does the test accurately measure what it intends to measure?	X			
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?		X		Listening#1 only has 5 questions even though it lasts 4: 08 min.
3. Are the items well distributed to make the test valid?	X			41 questions
4. Does test construction follow specific guidelines?	X			Assessment Handbook
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA		YES	NO	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			Radio show, postnuptial, Kyoto protocol.
2. Are the listenings level appropriate?	X			
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY** NA		YES	NO	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)		X		Listening #1 questions might be answered without listening to the audio.
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?	X			

Observations:

- Listening #2: instructions not so clear and layout is not well distributed (all items should be in the same page)

- Skills are not stated in each section of the exam.
- Listening #2 audio is incomplete.

Level Four Listening Assessment
Exam version A | **Reading exam**
Exam Date 2016-30 **Listening exam X**

I. **VALIDITY** NA	YES	NO		COMMENTS
1. Does the test accurately measure what it intends to measure?	X			
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			43 Questions
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			SPECS
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA	**YES**	**NO**		**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			Recycling, break ups, divorce.
2. Are the listenings level appropriate?	X			
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY** NA	**YES**	**NO**		**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)		X		Students can answer some questions by their prior knowledge.
4. Is the answer to each question text dependent (it does not depend on other stems and keys)		X		Some keys help answer other questions of the listening section.
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?		X		Listening #2. part B distractors are not that effective.
8. Are the numbers of questions in order?		X		(listening 1) Question #1 is missing.
9. Do matching exercises have two extra options?	X			

Observations:

- Listening #2 follows all SPECS guideline.

- Grammar and vocabulary exercises are the same as they were before Specs.

Level Four Listening Assessment
Exam version A | Reading exam
Exam Date 2017-02 Listening exam X

I. **VALIDITY** NA		YES	NO	**COMMENTS**
1. Does the test accurately measure what it intends to measure?	X			-
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			45 Questions
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			Test Specs
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA		**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			- The global change effect - Marriage bliss - Divorce in Japan
2. Are the listenings level appropriate?		X		LISTENING#1 IS 5:51. LISTENING #2, the speed is too slow.
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY** NA		**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise (it clearly indicates what students have to do)?	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?			X	

Observations:

- Listening #2 details does not show how many points it has.

Level Two Reading Assessment
Exam version A | **Reading exam X**
Exam Date 201530 **Listening exam**

I. **VALIDITY** NA		YES	NO	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		The test does not include the course outcome "differentiate between opinion and fact"
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?		X		25 questions. Not enough to measure skills.
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			Assessment handbook
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA		**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			"Breaking up" "Tying the knot" "Getting a divorce"
2. Are the readings level appropriate?	X			
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?		X		There are two exercises that ask students to respond according to what "they have learned", this expression can be ambiguous when choosing the right answer.
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY** NA		**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)		X		Students do not need to read the text to answer the question correctly.
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)		X		There are some keys

				that do not follow the same format. (Nouns, verbs, adjectives, etc.)
6. Are the questions formulated positively?	X			
7. Are distractors well designed?		X		There are no distractors
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?			X	

Level Two Reading Assessment
Exam version A | **Reading exam _X_**
Exam Date 201610 **Listening exam**

I. **VALIDITY** NA		**YES**	**NO**	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		The test does not have the outcome "differentiate between opinion and fact"
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			43 Questions
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			Assessment handbook
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA		**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			"New job?" "The fossil fuel controversy"
2. Are the readings level appropriate?	X			
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			Climate changing
III. **TEST ITEMS QUALITY** NA		**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?	X			Vocabulary exercises

Level Four Reading Assessment

Exam version ONE | **Reading exam _X_**
Exam Date 201630 **Listening exam**

I. **VALIDITY** NA		YES	NO	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		The test does not include the learning outcome "differentiate between opinion and fact"
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			49 questions
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			Test Specifications
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA		**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			"College start-ups" "Climate change controversy"
2. Are the readings level appropriate?	X			
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?	X			
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			Topics: Careers of the future Is our climate changing
III. **TEST ITEMS QUALITY** NA		**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)		X		Reading for details#1: Students do not have to go to the text to answer the questions. (segment only)
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?	X			

Level Two Reading Assessment
Exam version A | **Reading exam X**
Exam Date 201710 **Listening exam**

I. **VALIDITY** NA		YES	NO	**COMMENTS**
1. Does the test accurately measure what it intends to measure?		X		The outcome "differentiate between opinion and fact" is not presented in the test.
2. Does the test have reasonable number of questions that can be completed by students within the expected time frame?	X			31 questions
3. Are the items well distributed to make the test valid?	X			
4. Does test construction follow specific guidelines?	X			TEST SPECS
II. **TEXTS LANGUAGE LEVEL APPROPRIACY** NA		**YES**	**NO**	**COMMENTS**
1. Is the language in the test representative of real-world language use?	X			
2. Are the readings level appropriate?	X			
3. Does the test have items that are contextualized rather than isolated?	X			
4. Does the test contain items/tasks that are unambiguous to the test-taker?		X		Reading #2 Make inferences (question #10) is not very clear or it is not formulated properly.
5. Is the test developed with a specific purpose and a particular group of test-takers?	X			
III. **TEST ITEMS QUALITY** NA		**YES**	**NO**	**COMMENTS**
1. Is the stem clear and precise? (It clearly indicates the kind of answers that students need to give)	X			
2. Is each key clearly identified as the answer to the question asked?	X			
3. Is the answer to each question text dependent? (it does not depend on students' prior knowledge)	X			
4. Is the answer to each question text dependent (it does not depend on other stems and keys)	X			
5. Are the keys of the questions parallel? (formatting)	X			
6. Are the questions formulated positively?	X			
7. Are distractors well designed?	X			
8. Are the numbers of questions in order?	X			
9. Do matching exercises have two extra options?			X	

Observations:

- Read for main ideas does not ask main ideas questions accurately. (purpose, example of

Irony and another good title) in both Reading #1 and reading #2.

- Reading #2. Read for details (question #7) seems not to ask detailed questions but instead main ideas questions.
- Reading #2 Make inferences (question #10) is not very clear or it is not formulated properly.

Appendix D

General Findings

Listening

Validity	201510	201530	201610	201630	201710	FINDINGS
Level II	No outcome "organize information from multiple listenings" 37 questions True/False exercises should be allotted more points Assessment Handbook	No outcome "organize information from multiple listenings" Grammar and vocabulary sections 44 questions True/False exercises should be allotted more points Number of questions per section not even Assessment Handbook	No outcome "organize information from multiple listenings" Grammar and vocabulary sections 49 questions True/False exercises should be allotted more points Number of questions per section not even Assessment Handbook	No outcome "organize information from multiple listenings" 48 questions True/False exercises should be allotted more points Test Specs	No outcome "organize information from multiple listenings" 30 questions True/False exercises should be allotted more points Test Specs	None of the tests evaluates the outcome "organize information from multiple listenings" All tests have reasonable number of questions True/False exercises should be allotted more points All tests designed following specific guidelines Most tests have grammar and vocabulary sections Distribution of the items/questions is not even in two tests.
Level IV		42 questions Assessment Handbook	Number of questions is not coherent with the length of	43 questions Test Specs	45 questions Test Specs	All tests designed following specific guidelines Reasonable

			the listening (4'08- 5 questions) Assessment Handbook Items are not well distributed- not on the same page (instructions and layout) Strategies are not stated even though they are evaluated (main ideas, make inferences and details) Reasonable number of questions (41)			number of questions In one of the tests items are not well distributed and strategies are not mentioned
Text Language Level Appropriacy	**201510**	**201530**	**201610**	**201630**	**201710**	**FINDINGS**
Level II	Language is representative of real world use (Volunteer vacation, Assignment for a business class, }why people choose to	Language is representative of real world use (How games relate to real life, Description of the cities) Listening 2 too long (6'06'') The low	Language is representative of real world use (How games relate to real life, Food tasters talking about their	Language is representative of real world use (Assignment for a business class, Doctor's appointment) Listening 1	Language is representative of real world use (Doctor's opinion, High Tech health) One of the questions in make inference	Language is representative of real world use Listenings are too long Low speed of listenings In one exam, there is one listening not related to

	live where they do) Listenings 2 and 3 are too long (4’28’’, 4’51’’) The low speed might interrupt students’ attention Listening 3 is not related to the topics	speed might interrupt students’ attention Listening 3 is not related to the topics	job) Listening 2 too long (4’47’’) The low speed might interrupt students’ attention	too long (4’28’’) The low speed might interrupt students’ attention	section is ambiguous because it does not have a correct answer The low speed might interrupt students’ attention	the topics One question in the make inference section in one of the tests does not have a correct answer
Level IV		Language is representative of real world use (Responsibility, The great banana race, The envelope, Prenuptial agreement) Listenings 1 and 2 are too long (4’20’’, 5’13’’) Level appropriate, contextualized and unambiguous tests Specific purpose, particular group of test –takers, specific language use	Language is representative of real world use (Radio show, Post nuptial agreement, Kyoto protocol) Listenings 1 and 2 are too long (4’20’’, 5’13’’) Level appropriate, contextualized and unambiguous tests Specific purpose, particular group of test –takers, specific	Language is representative of real world use (Recycling, Break-ups, Divorce)		Language is representative of real world use Level appropriate Contextualized Unambiguous Specific purpose, particular group of test –takers, specific language use in mind

		in mind	language use in mind			
Test Items Quality	**201510**	**201530**	**201610**	**201630**	**201710**	**FINDINGS**
Level II	Listening 3 has a question that can be answered with informatio n of another question Matching exercise does not have extra options	All questions are text (audio) dependent Matching exercise has two extra options	All questions are text (audio) dependent Matching exercise has two extra options	All questions are text (audio) dependent Matching exercise has two extra options	Listening 1- Make inference – question 1 is not clear nor precise No matching exercise	Each key is clearly identified as answers to questions Keys of questions are parallel Questions are formulated positively Distractors are well designed Numbers of questions in order One questions is not text dependent No extra options in matching exercises in some tests
Level IV		Questions in listening 4 can be answered with students' prior knowledge Listening 2- some keys might give the answer to other questions Distractors	Questions in listening 1can be answered with students' prior knowledge	Some questions can be answered with students' prior knowledge Some keys might give the answer to other questions Distractors are not		Some questions can be answered with students' prior knowledge Some keys give the answer to other questions

		are not well designed (question 1- main ideas) No matching exercises		well designed (listening 2- part b) The number of questions is not in order (listening 1- questions 1 is missing)		

Reading

Validity	201510	201530	201610	201630	201710	FINDINGS
Level II	No outcome "differentiate between opinion and fact" 31 questions Assessment Handbook	No outcome "differentiate between opinion and fact" 47 questions Assessment Handbook Grammar and Vocabulary exercises	No outcome "differentiate between opinion and fact" 44 questions Assessment Handbook No distribution of items.	No outcome "differentiate between opinion and fact" 44 questions Assessment. Test Specs Recycle text (Urban Farmer)	No outcome "differentiate between opinion and fact" 44 questions Assessment. Test Specs Recycle text (Urban Farmer) (201530-201610) Reading for purpose is not an outcome of the level.	**Both before/ after** No outcome "differentiate between opinion and fact " Reasonable number of questions. **Before:** Assessment handbook 201530 (grammar/voc exercise) 201610 (no distributions of items) **After:** Both (201630/201710) Test specs Recycled text questions 201710: reading for purpose is not an outcome It does not evaluate information from multiple texts.
Level IV		No outcome "differentiate between opinion and fact" 25questio	No outcome "differentiate between opinion and fact" Same	No outcome "differentiate between opinion and fact" 49	No outcome "differentiate between opinion and fact" 31 questions Assessment Test Specs Reading 1 and 2	Before and after: No outcome "differentiate between opinion and fact " Questions are

		ns no reasonable # of questions. Assessment Handbook Grammar	outcome of level II 43 questions Assessment Handbook Grammar	questions Assessment Test Specs	main ideas are not accurately asked. (purpose, example of irony and another good title) Details questions are evaluated as they were main ideas.	well distributed. **Before:** Assessment handbook 201530 (no reasonable # of questions) **After:** Test Specs Reasonable # of questions Main ideas not properly asked Detailed questions evaluated as main ideas.
Text Language Level Appropriacy	**201510**	**201530**	**201610**	**201630**	**201710**	**FINDINGS**
Level II	Real world language use Reading# 1 too long divided	Real world language use (food, urban farmer)	Real world language use Language use in mind does not correspond with the language studied in class. Reading p6 questions p2	Real world language use Savant people, urban farmer, recycled reading (201530)	Real world language use Reading are not level appropriate (Present perfect-wording in the text might be more complex than the wording in the questions.) reading #1 question 4	**Before and after:** Real world language use Items and tasks are unambiguous Before: Handbook 201510 reading #2 too long 201610 does not correspond with the language studied in class.

						After: 201710 readings are not level appropriate
Level IV		Real world language use (breaking up, tying the knot, getting divorce) Some items/tasks are ambiguous (what have you learned)	Real world language use (new job, fossil fuel controversy)	Real world language use (college star-ups, climate change controversy)	Real world language use Ambiguous items and tasks(reading 2 make inference questions) 10 is not clear, not formulated properly	Before and after: Real world language use Level appropriate texts Contextualized items Specific purpose, language in mind Before: 201530 ambiguous items/tasks After: 201710 ambiguous items/tasks
Test Items Quality	**201510**	**201530**	**201610**	**201630**	**201710**	**FINDINGS**
Level II	Answers to questions are not text dependent	The stem is not clear and precise-reading 1 details Questions 3-4 don't indicate what to do No extra options in matching exercises	Complies with	Complies with	Reading 2 questions are the same (main ideas, details, make inferences) from 201630 test No matching exercises Obvious distractors(especially voc section) Two options are formulated negatively	Before: 2015 (no text dependent questions- no matching ex) 2016 (no clear stems) No extra options in matching ex After: The same recycled reading and questions No matching ex

						Obvious distractors Two questions are formulated negatively
Level IV		No text dependent questions (prior knowledge) No parallelism (nouns, verbs, adjs) No distractors.	Complies with	No text dependent questions (stems and options help answer correctly) reading #1 details	No matching exercises	Before: 201530/201630 No text dependent questions 201530 no parallelism / no distractors 201610 excellent 201710 no matching ex, the rest is ok.

Authors' Biography

Viviana Daniels Coneo holds a Bachelor's degree on Education in Foreign Languages Teaching from Universidad del Atlántico (2007) and a Postgraduate Diploma on English Language Teaching from Universidad del Norte (2012). She has been working as an English teacher for eleven years in different institutions in Barranquilla. She worked for four years at Centro Cultural Colombo Americano teaching mostly children and adults. After that she worked in different private and public schools with students in elementary and high school. In 2010, she started working at universities, first at Universidad Simon Bolivar coordinating levels five and six of the English Language Program and teaching students from different majors. At Universidad del Norte, she started coordinating the program English For Schools and then she worked in a project providing academic assistance to English teachers in a public school (Normal de Manatí) in terms of revision of curriculum, lesson planning and assessment. Currently, she works in the Undergraduate program with the credit-bearing program, teaching students from the programs of nursing, business administration, accounting, among others.

Elkin Jose Villanueva Niebles holds a Bachelor's degree on Education in Foreign Languages teaching from Universidad del Atlántico (2012), and a Postgraduate Diploma on Teaching English as a foreign language from Universidad del Norte (2015). The needs analysis was about how to improve reading skills by the implementation of reading strategies. Elkin has been working as a language English teacher for six years. He worked for five years at Centro Cultural Colombo Americano (CCCA) teaching children, teenagers and adults. He worked in different schools that had agreement with CCCA such as Colegio Lourdes, Colegio Colón, Colegio Buen consejo, Colegio Eucarístico, Colegio 20 de Julio, Colegio Domingo Sabio, Colegio Las

Mercedes, Colegio San Roque and the Company Sociedad Portuaria de Barranquilla. In this period, he was the coordinator of Colegio Eucarístico and Sociedad Portuaria de Barranquilla for one and a half year. He is currently working as an English teacher at Universidad del Norte in the Undergraduate program with credit-bearing program, teaching students from Accounting, Business Administration, Philosophy and Nursing. The levels he has taught at this Universidad del Norte are Nivelatorio, English I, English III, English VII, Business English VII, and English VIII. He has been working at this university since January 23rd 2016.